
Cell
Volume 163
Number 3
October 22, 2015
www.cell.com

In This Issue

Cell

Reading within the LINEs

DENLI ET AL., PAGE 583

A primate-specific open reading frame in LINE-1 retrotransposons both enhances LINE-1 mobility and promotes generation of exon fusion proteins, contributing to retrotransposon-mediated diversity.

The Making of a Mass Murderer

RASMUSSEN ET AL., PAGE 571

The plague-causing bacteria *Yersinia pestis* infected humans in Bronze Age Eurasia three millenia before any historical records of plague but only acquired the genetic changes making it a highly virulent, flea-borne bubonic strain about 3,000 years ago.

Evolutionary Potential of Promiscuity

AAKRE ET AL., PAGE 594

Interacting proteins can coevolve through the generation of promiscuous variants, which serve as mutational intermediates that preserve the ability of the two proteins to functionally interact while they evolve.

EF-fective Intoxication

WHITNEY ET AL., PAGE 607

The ubiquitous translation factor EF-Tu steps out of its typical role to help a *Pseudomonas* toxin access a target cell, which is then sent into stasis rather than being killed.

Disordered Conduct in Context

FREDERICK ET AL., PAGE 620

Sensitivity-enhanced NMR enables structural analysis of proteins at endogenous levels in a native biological context and reveals that the cellular environment alters the structure of an intrinsically disordered protein domain.

Unzipping Recognition

RUBINSTEIN ET AL., PAGE 629

Protocadherin isoforms mediate neuronal self-recognition through a zipper-like association mechanism that allows recognition of isoform mismatches and chain-termination of the interactions.

The Futility of Staying Warm

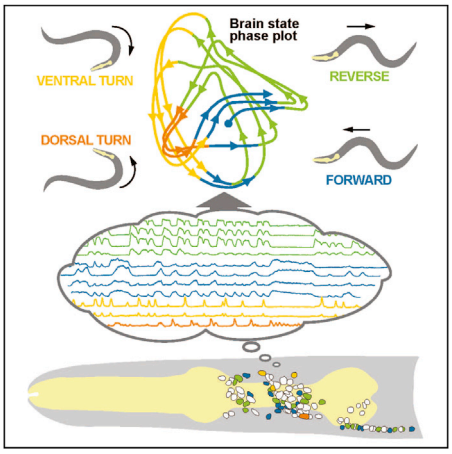
KAZAK ET AL., PAGE 643

Beige fat uses a futile cycle of creatine phosphorylation to dissipate energy and stimulate mitochondrial ATP demand, thereby promoting thermogenic cold adaptation.

Collective Brain Dynamics Code Behavior

KATO ET AL., PAGE 656

Simultaneously recording the activity of nearly all neurons in the *C. elegans* brain reveals that most active neurons share information by engaging in coordinated, dynamic network activity that corresponds to the sequential assembly of motor commands.



Tails of Ethylene Signaling

LI ET AL., PAGE 670
MERCHANTE ET AL., PAGE 684

A translational repression mechanism in which the 3′ UTRs of mRNAs act as signal transducers that relay ethylene signaling in plants.

Synthesizing Splicing Predictions

ROSENBERG ET AL., PAGE 698

A combination of synthetic biology and machine learning enables identification of universal patterns of RNA splicing based on sequence motifs and prediction of the effects of disease-related human SNPs.

Hold On Loosely

HEIN ET AL., PAGE 712

Weak interactions shape the cellular protein interaction network, as determined from proteomic measures of interaction specificities and strengths and protein copy numbers.

CFTR's Opening Wave

SORUM ET AL., PAGE 724

As the CFTR channel pore opens, a "wave" of conformational changes propagates through the protein complex, straining the interface where the prevalent cystic fibrosis mutation occurs.

Fast and Flexible

MILLES ET AL., PAGE 734

Intrinsically disordered nucleoporins engage nuclear transport receptors through many minimal, weakly binding motifs to form polyvalent complexes that retain conformational plasticity, thus ensuring both rapid and specific transport.

Sweetening the Deal on an Anti-Viral

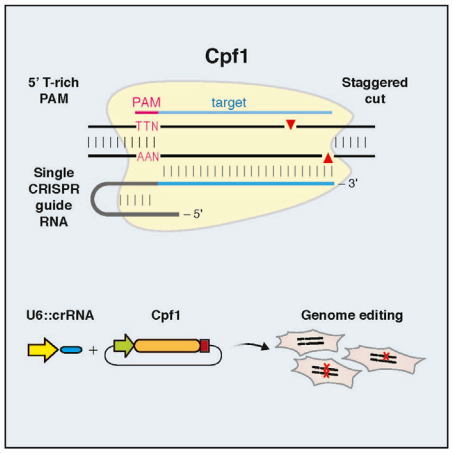
SWANSON ET AL., PAGE 746

Eliminating the mitogenic activity of a lectin by recalibrating how the protein "reads" surface carbohydrates expands its therapeutic potential as a broad-spectrum anti-viral agent.

Move over Cas?

ZETSCHE ET AL., PAGE 759

Cpf1 is an RNA-guided DNA nuclease that provides immunity in bacteria and can be adapted for genome editing in mammalian cells.



Cell

William Erwin Paul (1936-2015)

Legacy is defined as something handed down from the past, as from an ancestor or predecessor. The legacy of a great scientist includes not only their seminal work, but also the people they have mentored and inspired. Immunologist William Erwin Paul, who died on September 18 of this year, left just such a legacy.

Bill Paul, as everyone knew him, was an extraordinary scientist, administrator, and mentor. He handed down a major body of scientific work, an internationally renowned department at the National Institutes of Health (the Laboratory of Immunology, which he headed for 45 years), and many dozens of trainees who are still grateful for his mentorship and the opportunities that he gave them. As two of those mentees, we feel privileged to help share his story and the example he set for other scientists. Especially in an era that seems to reward shameless self-promotion and a disdain for teaching and administrative activities, Bill's life and work across this whole spectrum stands out in marked contrast to the self-absorption that some feel is necessary for a successful career in science.

Born in Brooklyn in 1936, Bill's parents were Jack Paul, an immigrant from the Ukraine, who ran an auto body shop, and Sylvia Gleicher, who came from a family of scientists. He first became interested in immunology by reading a collection of essays by Michael Heidelberger on a Brooklyn trolley, and following an internship and residency at what is now the Boston Medical Center and the National Cancer Institute in Bethesda, MD, he found his way back to New York and into the Lab of Baruch Benacerraf at NYU. When Benacerraf accepted a position as head of the Laboratory of Immunology (LI) at NIAID in Bethesda, Bill moved with him, and he remained at NIH the rest of his life. Within a year, Benacerraf was recruited to Harvard and Bill became chief of the laboratory at the tender age of 34 in 1970.

At the NIH, there are different styles with which a lab chief can manage his or her charges, with some choosing to view everyone as an extension of their own research agenda. But here, as in many other ways, Bill set himself apart by allowing each group leader the freedom to pursue his or her own ideas while he shouldered the administrative burden for 45 years. It's hard to imagine anyone doing that these days.

All the while, he ran a group of his own and had many singular scientific accomplishments—first in cellular immunology, where he and his colleagues discovered IL-4, a key cytokine involved in T cell activity, particularly in allergic and inflammatory diseases. This became the major focus of his lab, which he completely converted to molecular biology in order to clone and characterize the IL-4 gene and its transcriptional regulation.

But another opportunity to serve in an important administrative role arose in 1993, when the NIH came under great pressure from AIDS activist groups to move more quickly to develop an effective treatment. In response, the decision was made to create an Office of AIDS Research at the NIH in order to coordinate HIV research. Here, Bill's calming presence and deeply analytical skills were put to great effect, earning the trust of the activist community and helping to organize the many scientific moving parts to be more effective. Out of this effort came the amazingly effective anti-HIV therapies that are used by more than 15 million people today, transforming what was a death sentence into a chronic disease. Bill was also very proud that, during this period, he was able to directly help persuade President Clinton to establish the Vaccine Research Center at NIH, which has been an important force in advancing both vaccine research and candidate vaccines into the clinic.

Somehow, he found the time to participate in a myriad of national and international service activities, serving on prize committees, and review panels and as President of the American Association of Clinical Investigators and the American Association of Immunologists. He also developed an advanced textbook (*Fundamental Immunology*), published in 1984 and carried through seven editions. It has been rightly called *the* advanced text for any serious student of the subject. Clearly, he had a passion for the subject and a passion to communicate that enthusiasm in almost every possible way.

On a personal level, he was everyone's dream mentor—immensely thoughtful and knowledgeable, kind, and optimistic, but also very rigorous. He gave those who worked with him the freedom and encouragement to do their best. In one of our cases (M.M.D.), we came to the LI largely ignorant of cellular immunology, which at that time was largely separate from the molecular version (so separate that Niels Jerne referred to "cis" and "trans" immunologists), but with cloning skills that were rare at that time and the idea to use said skills to investigate gene expression differences between B and T lymphocytes.

William Erwin Paul (photo by Alena Soboleva)

Bill, while professing to have no expertise to offer this unusual endeavor, nevertheless asked very astute questions and also (M.M.D. later learned) queried some noted molecular immunologists as to whether this was advisable. Apparently it was, because he gave it his go-ahead and, as the results started to look promising, steered people looking for interesting projects toward it, leading to a small but effective branch of the lab that, with the addition of Steve Hedrick, endeavored to clone the genes behind the near legendary T cell receptor—key to understanding T cell specificity, the lack of which was a major stumbling block in the field at that time. All through this, Bill offered constant encouragement and those very good questions. But most remarkable of all, when that effort was actually successful despite intense competition—akin to winning a major lottery—Bill refused to take any credit, even though M.M.D. was all the while a postdoc under his tutelage. But this was who he was: a scientist's scientist and an example to us all.

L.H.G. came to Bill's lab in 1979 as an MD postdoctoral fellow with only a year of research experience, needing all of the training and mentoring she could get. In Bill, she found someone who believed in providing lots of both. Since she was also 8 months pregnant, Bill must have wondered exactly how much work he was going to get out of her. When she returned to work a week after her daughter was born, perhaps he breathed a sigh of relief.

In science, there are relatively safe projects and then there are somewhat risky projects. But Bill suggested that she take on an extremely risky project that no one, probably including Bill, thought could possibly work. But he was an incredible optimist—one of the things that we loved most about him. So L.H.G. plunged ahead and, to everyone's astonishment, succeeded in generating mutant class II MHC antigen-presenting cells. The results turned out to be quite important in understanding the structure-function relationship between the T cell receptor and its MHC ligands.

Thinking back on this, practically everything Bill did turned out to be important. He had a great nose—or a green thumb, if you will—for sniffing out what would matter. For him, if a project didn't reach for the moon, or at least the stars, then it wasn't worth doing. This important lesson has inspired us throughout our careers to dare to take big risks where the payoff for science would be huge even though the likelihood of success might be small.

Bill was also one of those rare individuals at the time who always supported women. He had many female postdocs, who called themselves Bill's lymphettes. When L.H.G. insisted on returning to the lab the day after she delivered her second child just 2 years later, Bill didn't blink an eye. Not many mentors would have borne with equanimity the arrival of a postdoc who was 8 months pregnant and who then proceeded to have a second child. Many young women nowadays might find it hard to believe how little understanding professional women could receive in those days when they tried to combine career and family. In that regard, as in many others, Bill was a revolutionary. It may have helped that Marilyn Paul, Bill's beloved and amazing wife, was herself a very strong and also supportive personality.

In conclusion, it's hard for something as impersonal as words to give a sense of a life as richly lived, as deeply beneficial to so many people, as the life of Bill Paul. Bill's legacy was not only his great contributions to our understanding of the way the immune system works, but also his immeasurably important contributions to the scientific and personal development of the many people who had the good fortune to work with him. The great English poet Robert Browning once said that the best measure of the height of a man is the length of the shadow his mind casts. By this measure, Bill Paul was a giant.

Mark M. Davis[1,*]
and Laurie H. Glimcher[2,*]
[1]Howard Hughes Medical Institute, Department of Microbiology and Immunology, Institute of Immunity, Transplantation, and Infection, Stanford University School of Medicine, Stanford, CA 94304, USA
[2]Office of the Dean and Department of Medicine, Weill Cornell Medicine, New York, NY 10065, USA
*Correspondence: mmdavis@stanford.edu (M.M.D.), lglimche@med.cornell.edu (L.H.G.)
http://dx.doi.org/10.1016/j.cell.2015.10.024

Cell

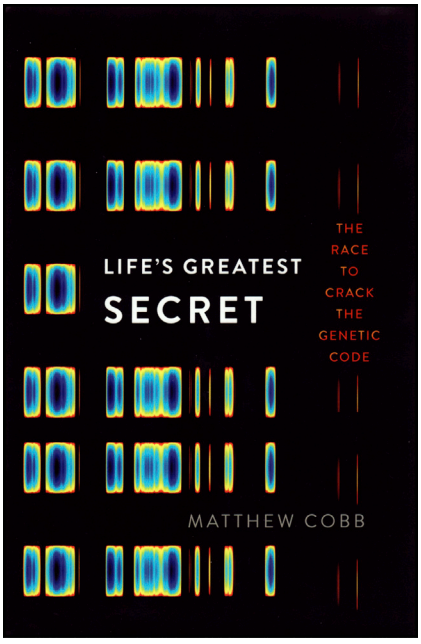
Unveiling the Code of Life

Life's Greatest Secret: The Race to Crack the Genetic Code
Author: Matthew Cobb
Basic Books: New York, NY, USA (2015). 464 pp. $29.99

There is no shortage of written contributions on various aspects of the history of genetics and molecular biology—*The Eighth Day of Creation: The Makers of the Revolution in Biology* stands out as a premier example, memorably focusing on reaching the intelligent lay reader. Matthew Cobb's contribution is unquestionably written for scientists. But it too deserves adulation as a masterwork.

The introductory chapter *Genes before DNA* reminds readers of the familiar early pioneers of genetics, including Gregor Mendel, Hugo de Vries, Theodor Boveri, Wilhelm Johannsen, Thomas Hunt Morgan, and Nikolai Koltsov. This chapter also thoughtfully informs us of the important intellectual contributions of the eminent physicist Erwin Schrodinger, who is credited with the first notion of a "code script" when talking about how genes operate. The succeeding chapter called *Information Is Everywhere* may tempt all but the most intellectually oriented readers to toss the book aside. My advice is to curb this impulse should it arise!

This reviewer was particularly entranced by Cobb's treatment of the famous transformation experiments executed by Oswald Avery and his colleagues Colin MacLeod and Macleod McCarthy in the mid-1940s that led them to the conclusion that genetic information resides in DNA rather than proteins, the latter being the alternative and widely held view in the genetics community. The erudition of this chapter lies in an element that particularly distinguishes Cobb's writing, namely the historical depth that he has brought to this literary contribution. Most, if not all, students are taught that Avery was the first to experimentally demonstrate that genes are made of DNA. But few are likely aware of the enormous challenges that he had to endure from the unshakable adherence to the entrenched notion that genes are made of protein and that, if DNA was in anyway involved in gene action, it was surely by way of some subsidiary (perhaps structural) role. Besides, DNA was then considered an utterly boring molecule equipped with just the four bases, deoxyribose and phosphate, hardly persuasive "to bring about the almost infinitely different effects produced by genes." Cobb informs that both the experiments of Avery and his group and the equally famous later experiments of Alfred Hershey and Martha Chase were persistently dogged by a criticism that was essentially impossible to definitively address, namely that they could never definitively prove that their transforming principle contained absolutely no protein. Cobb reveals that the influential biologist Alfred Mirsky, together with Arthur Pollister, published a widely read article that stressed that "there can be little doubt in the mind of anyone who has prepared nucleic acids that traces of protein probably remain in even the best preparations and that as much as 1 or 2 per cent of protein could be present in a preparation of pure, protein-free nucleic acid." Even the celebrated geneticist Herman Muller wrote in an article that he was personally convinced that Mirsky's suggestion that undetected "genetic proteins floating free in the medium caused Avery's results."



Cobb points out that, regardless of "the overwhelming evidence, all of which suggested that the transforming principle was made of DNA and that genes may be too," the final paragraph of the paper in which Avery and his colleagues announced their startling findings "opened with a phrase that suggested that the team was not quite as confident as they ought to have been." "It is of course possible that the biological activity of the substance described here is not an inherent property of the nucleic acid but is due to minute amounts of some other substance adsorbed to it or so intimately associated with it as to escape detection," Avery et al. wrote. But, they also boldly stated, "there is no evidence in favor of such a hypothesis that is chiefly supported by the traditional view that nucleic acids are devoid of biological specificity." Distressingly, this traditional view hung around in the minds of many scientists, even prominent ones, for years, causing Avery to suffer frank clinical depression. And when Avery died in 1955, "the brief obituary that appeared in the New York Times did not even mention DNA."

Regardless, over the years, Avery's contention stimulated the thinking and work of an increasing cadre of established and future stars in genetics, including Joshua Lederberg. Cobb notes that the journal *Nature* described Avery's work in glowing terms, and a (small) number of scientists were in fact highly complementary. In October, 1944, the New York Academy of Medicine awarded Avery its Gold Medal. And in 1945, the Royal Society of London graced his experimental achievements with the Copley Medal. But Avery was never graced with the highly deserved distinction of Nobel Laureate.

Cobb interrupts the progress of his history with another epistemological chapter dubbed *The Age of Control* in which he outlines the discipline of cybernetics, a term that it is relevant to the study of systems, including mechanical, physical, biological, cognitive, and social systems. Cybernetics is applicable when a system being analyzed incorporates a closed signaling loop—i.e., where action by the

system generates some change in its environment and that change is reflected in the system in some manner (feedback) that triggers a system change. The intent of this chapter is to alert the reader to the emergence of cybernetics when feedback mechanisms in molecular biology were discovered by later makers and shakers in molecular biology, notably from the exquisite experiments on gene regulation executed by the famous French duo of Francois Jacob and Jacques Monod described in a later chapter.

Much of the rest of the book covers the history on the elucidation of the structure of DNA (a topic well covered in James D. Watson's *The Double Helix*) and the pursuit of the Holy Grail—deciphering the genetic code. Cobb peppers his writing of the latter seminal breakthrough with delightfully interesting anecdotal information that displays the depth of research for his book. He informs the reader:

"On March 19, 1953, about two weeks after the double helix model had been completed, Francis Crick wrote a letter to his 12-year old son, Michael, who was at boarding school. Crick told Michael what he had discovered, and included a sketch of the structure of DNA. He then went on to explain the significance of the double helix. 'It's like a code,' Crick wrote to Michael. 'If you are given one set of letters you can write down the others. Now we believe that the D.N.A. is a code. That is, the order of the bases (the letters) makes one gene different from another gene (just as one page of print is different from another).'"

Cobb relates that, while the notion that the sequence of bases in a DNA chain had been speculated for some time, this letter to his very young son was the first time that anyone had stated in writing that DNA contains a code. In 2013, the letter fetched $6 million at an auction! Crick's leadership, intellectual genius, and scintillating personality during the period in which the genetic code was slowly but surely unraveled leap majestically from Cobb's pen.

Equally arresting and presumably little-known historical anecdotes surface when Cobb relates that, after Marshall Nirenberg (an unknown scientist to most of the molecular biology community) reported his use of homopolymers to elucidate the genetic code in a 10 min talk at the Fifth International Congress of Biochemistry in Moscow in August 1961, Matt Meselson informed Crick of these electrifying experimental results, prompting Crick to invite him to present his findings again in a longer plenary talk at a symposium that Crick was to chair the following day. Following his second presentation, Nirenberg was so gratified and elated he was prompted to comment:

"The reception was really remarkable, fantastic. I remember Matt Meselson, who was sitting right up front. I didn't know him at the time, but he was so overjoyed about hearing this stuff that he impulsively jumped up, grabbed my hand, and actually hugged me and congratulated me for doing that. I could have been part of a rock band or something. That really meant an awful lot to me. It really meant more to me than all kinds of awards and what-not because it was genuine and spontaneous."

The work in the Nirenberg laboratory and that of a competing laboratory led by the Spanish-born biochemist Severo Ochoa contributed mightily to deciphering the genetic code. In 1968, Nirenberg shared the Nobel prize in physiology or medicine with Robert Holley and Gobind Khorana.

Cobb concludes his book with a section entitled *Update* that details the history of molecular genetics and molecular biology to the present time, including the discovery of introns, the use of the polymerase chain reaction (PCR), sequencing entire genomes, paleogenomics, population genetics, evolutionary genetics, genetic engineering, the potential for synthetic biology, and more. The book also features a pleasing gallery of photos.

The dominance of individual brilliance in molecular genetics so remarkably displayed by Francis Crick and Sydney Brenner during the decades of 1950s and 1960s is fading all too rapidly. In his conclusion, Cobb addresses the perils of "big science," especially in the field of genomics, pointing to a paper in *Nature Genetics* published in 2014 that listed 440 authors! Cobb notes "It is now becoming commonplace, changing the relationship of individual scientists to the work they produce, rendering each person's contribution relatively minor and highly specific." This threatening shift in the sociology of science cries out for attention if molecular biology is to regain its former attraction to college students interested in pursuing careers in disciplines exemplified by modern day genomics.

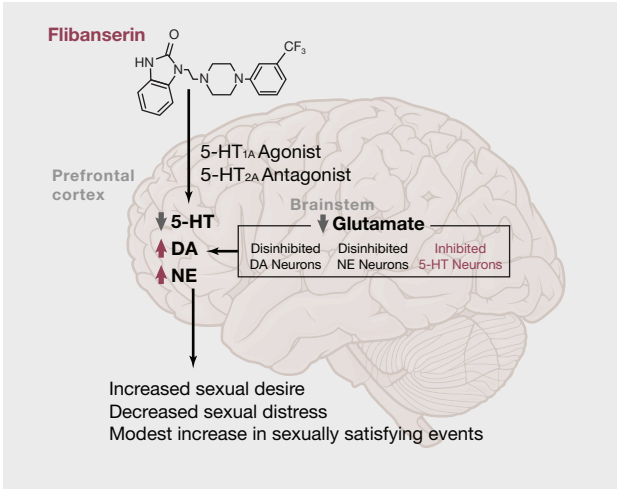
All in all, Matthew Cobb, who hails from the University of Manchester—which notably includes a Center for the History of Science, Technology and Medicine and whose eclectic historical contributions include the efforts of the French resistance during WWII—has presented the scientific and perhaps members of the non-scientific communities an erudite and comprehensive history that should be required reading for all graduate students in the disciplines of genetics and molecular biology and, most certainly, students of the history of science.

Errol C. Friedberg[1,*]
[1]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390-9072, USA
*Correspondence: errol.friedberg@utsouthwestern.edu
http://dx.doi.org/10.1016/j.cell.2015.10.018

Cell

Treatment for Hypoactive Sexual Desire

James G. Pfaus
Center for Studies in Behavioral Neurobiology, Department of Psychology, Concordia University, Montréal, QC H4B 1R6, Canada
Correspondence: jim.pfaus@concordia.ca
http://dx.doi.org/10.1016/j.cell.2015.10.015



Flibanserin acts at cortical, limbic, hypothalamic, and brainstem nuclei to inhibit serotonin release by binding to 5-HT1A autoreceptors and block postsynaptic action of serotonin at 5-HT2A receptors. This gradually disinhibits the turnover of other monoamines like dopamine and noradrenaline that are critical for sexual desire.

NAME
Flibanserin (Addyi)

APPROVED FOR
Treatment of hypoactive sexual desire disorder (HSDD) in women

TYPE
Small molecule: centrally active piperazine/benzimidazol derivative

MOLECULAR TARGETS
Full agonist at 5-HT1A receptors, antagonist at 5-HT2A receptors. Reduces forskolin-stimulated cAMP and eliminates 5-HT-stimulated phosphatidyl-inositol turnover in cortex.

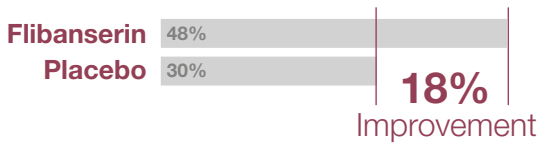
CELLULAR TARGETS
Reduction of serotonin-induced descending inhibition in medial prefrontal cortex, limbic regions, hypothalamus, and brainstem.

EFFECTS ON TARGETS
Disinhibition of dopamine (DA) and noradrenaline (NE) turnover within cortical, limbic, and hypothalamic regions associated with the stimulation of sexual desire.

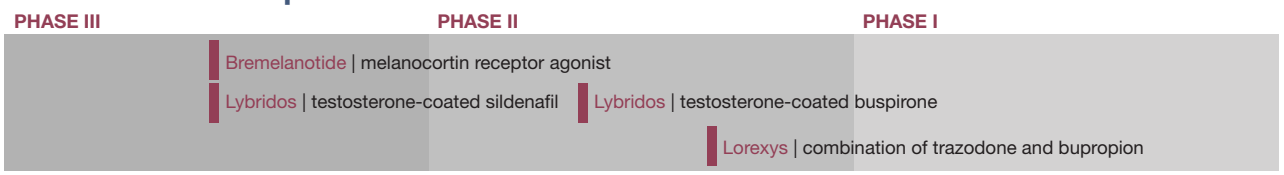
DEVELOPED BY
Boehringer Ingelheim > Sprout Pharmaceuticals > Valeant Pharmaceuticals

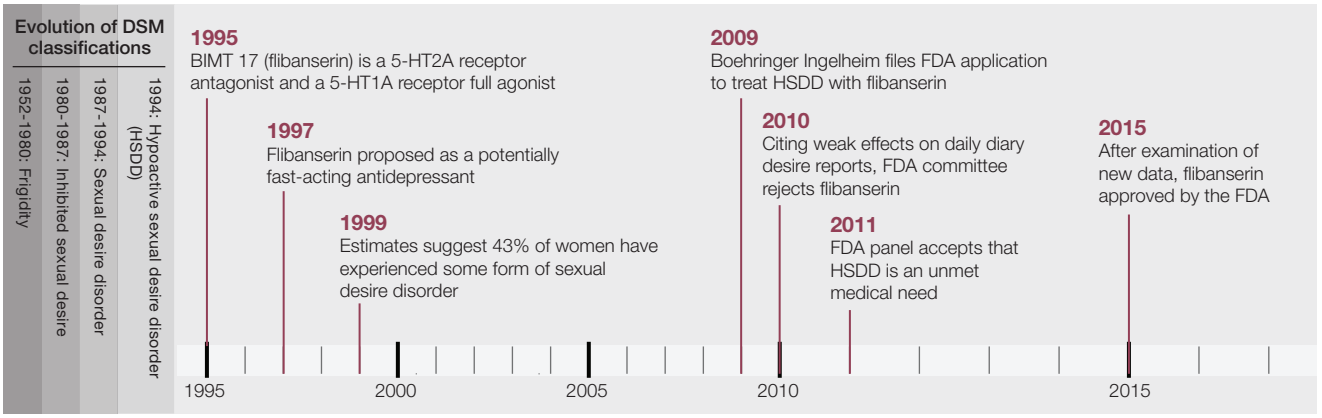
Varying estimates of HSDD incidence

After 24 weeks of treatment



Also under development for HSDD





References for further reading are available with this article online: www.cell.com/cell/abstract/S0092-8674(15)01329-X

Much Ado about Zero

Jef D. Boeke[1,*] and David Fenyo[1]
[1]Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Langone School of Medicine, New York, NY 10016, USA
*Correspondence: jef.boeke@nyumc.org
http://dx.doi.org/10.1016/j.cell.2015.10.033

LINE retrotransposons actively shape mammalian genomes. Denli et al. reveal a new open reading frame, ORF0, on the antisense strand of human LINE-1 encoding a small regulatory protein. This finding may represent the birth of an emerging retrotransposon gene that can adopt various fates, as it can be fused to adjacent host sequences.

Long interspersed repeats or LINEs are retrotransposons that have littered mammalian genomes since their divergence from other vertebrates hundreds of millions of years ago. The human version of this sequence, LINE-1, is active in germlines, early embryos, and the brain, as well as in selected human cancers (Goodier, 2014). LINEs are known as potent agents of genome instability by mobilizing themselves, other sequences that do not encode reverse transcription machinery, such as short interspersed repeats (SINEs), and a multitude of processed pseudogenes (Burns and Boeke, 2012). Two LINE-1 genes, ORF1 and ORF2, are encoded by the human LINE-1 sequence, and both are directly involved in retrotransposition (Moran et al., 1996; Figure 1). ORF1 encodes a nucleic acid binding protein that avidly binds single-stranded RNA in the ribonucleoprotein particle that serves as a retrotransposition intermediate, whereas ORF2 specifies a polyprotein with both endonuclease and reverse transcriptase activities. The ORF1 and ORF2 sequences were defined nearly 30 years ago, when the consensus sequence of the active subfamily of LINEs was deduced (Scott et al., 1987).

It therefore comes as a big surprise that this intensively studied element in fact sports a third open reading frame, dubbed "ORF0," within the 5′ UTR of the LINE-1 transcript and on the opposite strand as the ORF1 and ORF2 structural genes (Denli et al., 2015 [this issue of *Cell*]). How could something so obvious have been missed for so long? There are perhaps three major reasons. Unlike the two ORFs that we know so well, it is encoded on the antisense strand. Moreover, ORF0 is very short, encoding a 71 amino acid peptide, which is in marked contrast to ORFs 1 and 2 that collectively span nearly 5,000 bp. Finally, unlike ORFs 1 and 2, the ORF0 sequence is conserved only within the primate lineage, a strong argument that the sequence does not play a direct constitutive role for retrotransposition.

The super-short nature of ORF0 and overall lack of conservation rightly calls into question whether or not this is really a gene at all or just an accidental juxtaposition of codons. It is presumably a relatively newborn gene of the primate lineage, albeit one inhabiting the genome of a DNA parasite rather than that of the primates themselves. Denli et al. (2015) brought multiple lines of evidence forward to support that ORF0 is in fact functional. The LINE-1 sequence contains two promoters, the best known of which initiates at the first base pair of the element. It is the promoter responsible for expression of ORFs1 and 2 and serves as the template for retrotransposition. A second antisense promoter drives expression out of the left end of the element, and it has been adopted as a promoter by multiple human genes (Mätlick et al., 2006; Figure 1). ORF0 is well positioned to have its expression driven by this antisense promoter. Moreover, insertion of reporter genes and tags in frame with ORF0 in an otherwise intact and unremarkable LINE-1 element led to gene expression in embryonic stem cells, and mutation of the ORF0 AUG initiator codon eliminated such expression. In addition, the GFP-ORF0 fusion protein was localized to the nucleus. Interestingly, ORF0 protein encompasses one or two splice donor sites previously observed to be fused to splice acceptors inside or, more commonly, outside various copies of the LINE-1 element. Capped and ribosome occupied ORF0 transcripts were readily identified and were far more abundant in stem cells than in fibroblasts, as is also the case for full-length LINE-1 ORF1-2 transcripts. Further ribosome footprinting and RNA-seq analyses identified fusion transcripts between ORF0 and at least five human genes. In addition, phylogenetic analyses showed that

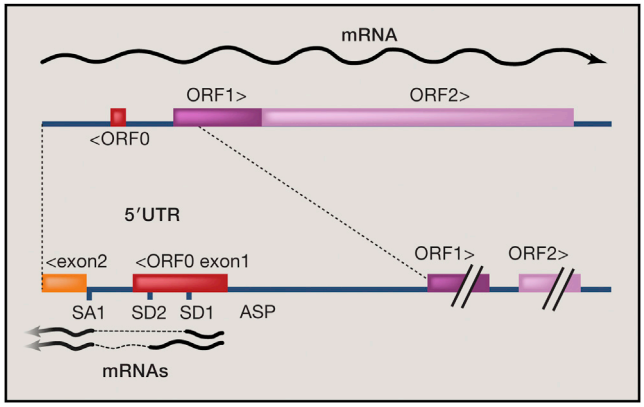


Figure 1. Relationship between ORF0 and Other Key Elements in Human LINE-1 Retrotransposon
The schematic (roughly to scale) shows the 5′ UTR region containing two promoters, ORF0, its downstream exon and signals for multiple splice-isoforms. Notably, the size of ORF0 is remarkably small, and it can be joined to a downstream exon to produce fusion protein product.

ORF0 could be reliably detected in ~50 copies in old world monkeys and thousands of copies in humans and great apes, but not in new world monkeys.

A critical question is whether ORF0 protein can be detected in non-engineered primate cells. Denli et al. (2015) provided evidence for the existence of the ORF0 protein using a combination of immunoprecipitation and mass spectrometry (MS). They overcame the issue of the mismatch between low ORF0 protein concentration and the limited dynamic range and sensitivity of MS by using polyclonal antibodies to enrich ORF0 protein. A second issue often encountered in MS analysis of short proteins is that, after digestion, there are often very few if any peptides amenable to MS sequencing, which need to be of a just-right length and well fragmented so that their sequences can be determined with high confidence. Denli et al. (2015) were able to obtain extensive fragmentation information almost entirely covering three tryptic peptides corresponding to ORF0 and its second exon (Figure 1). The MS detection was carried out on both overexpressed ORF0 protein and endogenous protein produced in human cells.

Just because a sequence is expressed does not make it a gene that encodes a functional protein. In this study, Denli et al. (2015) produced evidence suggesting a regulatory role for ORF0-encoded protein. Previous work had shown that an element driven by a promoter completely lacking LINE-1 sequences was active in retrotransposition, arguing strongly against a required role in *cis*. However, such a function might be provided in *trans*. Indeed, Denli et al. (2015) used a CAG-LINE-1 retrotransposition reporter element similar to those described earlier (Moran et al., 1996) to evaluate hopping frequency and showed that overexpression of ORF0 from a separate plasmid enhanced retrotransposition frequency by 41%. Thus, it seems likely that ORF0 plays some positive regulatory role in the retrotransposition process. It remains to be determined whether such a role of ORF0 is in any way related to its capacity in generating fusion protein containing host genomic sequences. Moreover, it would be interesting to see whether, and if so how, the ORF0 protein might functionally contribute to LINE-1 retrotransposition mechanistically.

REFERENCES

Burns, K.H., and Boeke, J.D. (2012). Cell *149*, 740–752.

Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C.N., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J., et al. (2015). Cell *163*, this issue, 583–593.

Goodier, J.L. (2014). Mob. DNA *5*, 11.

Mätlick, K., Redik, K., and Speek, M. (2006). J. Biomed. Biotechnol. *2006*, 1–16.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). Cell *87*, 917–927.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Genomics *1*, 113–125.

Evolutionary Reprograming of Protein-Protein Interaction Specificity

Eyal Akiva[1] and Patricia C. Babbitt[1,2,*]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

*Correspondence: babbitt@cgl.ucsf.edu

http://dx.doi.org/10.1016/j.cell.2015.10.010

Using mutation libraries and deep sequencing, Aakre et al. study the evolution of protein-protein interactions using a toxin-antitoxin model. The results indicate probable trajectories via "intermediate" proteins that are promiscuous, thus avoiding transitions via non-interactions. These results extend observations about other biological interactions and enzyme evolution, suggesting broadly general principles.

HEAD HEAL TEAL TELL TALL TAIL. This word game devised by Lewis Carroll requires moving from one word to another while keeping all intermediate words meaningful. It offers a nice analogy for a protein evolution model, where words represent functional proteins and mutations are word-to-word moves (Smith, 1970). It also represents one side of a debate, whether mutational navigation in sequence space from one protein function to another traverses via evolutionary intermediates that retain some functional features along the pathway to a new function. Because the evolution of new specificities in protein-protein interactions requires changes in at least two partners, the challenges for retaining functions that are vital for cell survival while evolving new ones may be more constrained (and more complicated) than in other

ORF0 could be reliably detected in ~50 copies in old world monkeys and thousands of copies in humans and great apes, but not in new world monkeys.

A critical question is whether ORF0 protein can be detected in non-engineered primate cells. Denli et al. (2015) provided evidence for the existence of the ORF0 protein using a combination of immunoprecipitation and mass spectrometry (MS). They overcame the issue of the mismatch between low ORF0 protein concentration and the limited dynamic range and sensitivity of MS by using polyclonal antibodies to enrich ORF0 protein. A second issue often encountered in MS analysis of short proteins is that, after digestion, there are often very few if any peptides amenable to MS sequencing, which need to be of a just-right length and well fragmented so that their sequences can be determined with high confidence. Denli et al. (2015) were able to obtain extensive fragmentation information almost entirely covering three tryptic peptides corresponding to ORF0 and its second exon (Figure 1). The MS detection was carried out on both overexpressed ORF0 protein and endogenous protein produced in human cells.

Just because a sequence is expressed does not make it a gene that encodes a functional protein. In this study, Denli et al. (2015) produced evidence suggesting a regulatory role for ORF0-encoded protein. Previous work had shown that an element driven by a promoter completely lacking LINE-1 sequences was active in retrotransposition, arguing strongly against a required role in *cis*. However, such a function might be provided in *trans*. Indeed, Denli et al. (2015) used a CAG-LINE-1 retrotransposition reporter element similar to those described earlier (Moran et al., 1996) to evaluate hopping frequency and showed that overexpression of ORF0 from a separate plasmid enhanced retrotransposition frequency by 41%. Thus, it seems likely that ORF0 plays some positive regulatory role in the retrotransposition process. It remains to be determined whether such a role of ORF0 is in any way related to its capacity in generating fusion protein containing host genomic sequences. Moreover, it would be interesting to see whether, and if so how, the ORF0 protein might functionally contribute to LINE-1 retrotransposition mechanistically.

REFERENCES

Burns, K.H., and Boeke, J.D. (2012). Cell *149*, 740–752.

Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C.N., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J., et al. (2015). Cell *163*, this issue, 583–593.

Goodier, J.L. (2014). Mob. DNA *5*, 11.

Mätlick, K., Redik, K., and Speek, M. (2006). J. Biomed. Biotechnol. *2006*, 1–16.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). Cell *87*, 917–927.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Genomics *1*, 113–125.

Evolutionary Reprograming of Protein-Protein Interaction Specificity

Eyal Akiva[1] and Patricia C. Babbitt[1,2,*]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

*Correspondence: babbitt@cgl.ucsf.edu

http://dx.doi.org/10.1016/j.cell.2015.10.010

Using mutation libraries and deep sequencing, Aakre et al. study the evolution of protein-protein interactions using a toxin-antitoxin model. The results indicate probable trajectories via "intermediate" proteins that are promiscuous, thus avoiding transitions via non-interactions. These results extend observations about other biological interactions and enzyme evolution, suggesting broadly general principles.

HEAD HEAL TEAL TELL TALL TAIL. This word game devised by Lewis Carroll requires moving from one word to another while keeping all intermediate words meaningful. It offers a nice analogy for a protein evolution model, where words represent functional proteins and mutations are word-to-word moves (Smith, 1970). It also represents one side of a debate, whether mutational navigation in sequence space from one protein function to another traverses via evolutionary intermediates that retain some functional features along the pathway to a new function. Because the evolution of new specificities in protein-protein interactions requires changes in at least two partners, the challenges for retaining functions that are vital for cell survival while evolving new ones may be more constrained (and more complicated) than in other

systems. How is cross-reaction between the evolving, homologous interaction partners evaded? What mutational trajectories do partners traverse while avoiding intermediate steps that may have negative biological consequences? In this issue of *Cell*, Aakre et al. (2015) utilize a model of toxin-antitoxin (TA) protein interactions that are essential for bacterial survival to study the problem systematically. Their results provide evidence for the preference for evolutionary paths involving biologically functional promiscuous intermediate steps, rather than switch-like trajectories that include non-interacting intermediates.

Bacteria typically include several chromosomally encoded paralogs of TA pairs in which an antitoxin neutralizes the toxin by interacting with it. Aakre et al. focus on the discrete problem of emergence of a new TA pair from an existing one. The specific ParD3/ParE3 interaction pair they chose for these experiments exhibits systemwide mutual exclusiveness, with almost no cross-reaction with other TA pairs that could complicate retrieval and interpretation of results.

Aided by a high-resolution crystal structure they solved for a specific ParD3/ParE3 complex, a 4-residue motif sufficient for reprogramming interaction specificity was identified. A mutation library was then constructed in which the one invariant Trp was retained while the three co-evolving positions of the motif were varied using only residues often found in natural ParD homologs. Fitness was approximated for intermediate stages in the path between one specific TA pair and another of different specificity using a competitive growth assay that allowed recovery of successful variants enriched over the time course of the experiments. For the ParD3 library, 252 variants were recovered that could effectively antagonize ParE3. As expected, repeating the competitive assay with another toxin, ParE2, produced a different set of 151 variants that neutralize this second toxin.

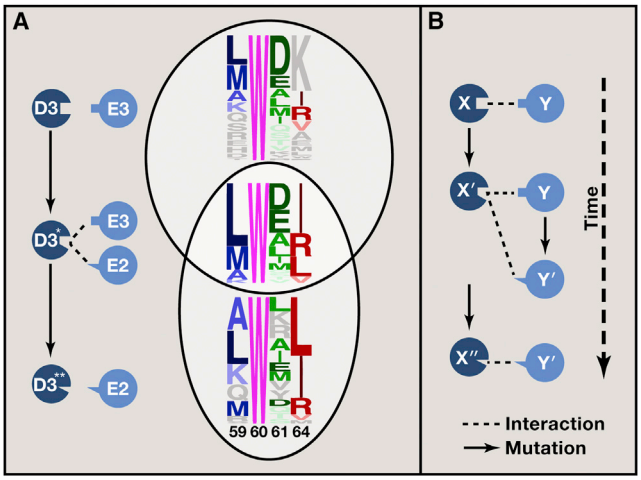


Figure 1. Reprogramming Specificity via Promiscuous Intermediates

(A) A Venn diagram showing the overlap between three sample sets of PraD3 antitoxin variants. Sequence logos represent the diversity in four specificity-determining positions that are overrepresented in ParD3 antitoxin variants that fit either ParE3 (the native toxin, upper), ParE2 (another toxin, lower) or both (middle). Amino-acid colors differ for each position; the darker the color, the more prevalent it is in the promiscuous motif. Grey residues represent cases in which the E3-specific logo or E2-specific logo includes residues that do not appear in the corresponding position in the promiscuous logo.

(B) The model suggested by Aakre et al. for reprograming protein-protein interaction specificity. An enabling, promiscuity-exerting mutation of protein X (X to X') allows protein Y to change its specificity determinant (Y to Y') and still bind both Xs. Protein X' then mutates (X' to X'') with an increase in specificity toward Y'. The protein-protein interactions are thus maintained throughout the evolutionary process. Features of this figure were adapted from Aakre et al. (2015).

The two antitoxin specificities are typified by distinct motifs in ParD3 specificity determinants (Figure 1A). While position 59 appears largely diffident in its variation pattern, position 61 is enriched for negatively charged residues for ParE3-specific variants, in contrast to small hydrophobic/positively charged residues in the ParE2-specific variants. Similarly, position 64 is enriched for positively charged residues in ParE3-specific variants, compared to small hydrophobic residues in ParE2 specific variants. Importantly, 31 variants exhibit dual-specificity toward ParE2/3, characterized by ParE3-like specificity at position 61 and ParE2-like specificity at position 64. Strikingly, evaluation of all alternative mutational trajectories between the two distinct specificities sampled shows statistically significant overrepresentation of traversal via promiscuous intermediates. Mutational trajectories also show significant enrichment for epistasis, rather than additive effect of mutations, consistent with similar findings in the evolution of enzyme-substrate interactions (Weinreich et al., 2006).

To investigate the important question involving co-evolution in interacting proteins, the authors performed another experiment traversing the sequence space from ParD3/ParE3 to ParD3*/ParE3*. Again, they found a prevalence in intermediate promiscuous variants, and, most importantly, that *all* presumed trajectories traversed via at least one promiscuous intermediate, suggesting the plausibility of this evolutionary path. Figure 1B summarizes these results, in which an X-Y interaction evolves to the orthologous X''-Y' interaction in at least three steps: (1) Mutation(s) in X to X' broadens specificity, allowing (2) Y to form a mutant, Y', that has the potential to interact with X as well as X', and finally (3) X' is mutated to X,'' narrowing its specificity to Y'.

Although reconstituting a natural history of protein repurposing is challenging and cannot be explicitly determined, this work contributes important initial observations toward this goal. Typically, mutation libraries sample a fraction of the sequence space. The approach used in this work allowed exclusion of infrequent, albeit viable trajectories, by focusing on the four most relevant positions and targeting only residues commonly appearing in contemporary proteins. Epistatic constraints and the occurrence of intermediates of modified or reduced function have been demonstrated for other types of models including in enzyme evolution, (Aharoni et al., 2005) and receptor-ligand evolution (Ortlund et al., 2007), and its practical implications have been exploited for protein-protein interaction engineering (Kortemme et al., 2004). Placed in this broader context, Aakre et al. provide new evidence for extending these conjectures to protein-protein interactions.

This work also suggests avenues for future research. For example, the

contributions of neutral drift and the impact of a few large-effect mutations versus many small-effect ones will need to be evaluated. Work on protein-protein interactions should be extended to other systems where cross-reaction is an issue, such as in other TA modules documented as cross-reacting (Zhu et al., 2010). Cross-reaction is also pertinent in other types of natural systems, as has been shown in some SH3 systems (Zarrinpar et al., 2003) and in the evolution of metabolic pathways (Kim and Copley, 2012). Ultimately, there are many other variables likely to be relevant in natural evolution that will surely be more difficult to ascertain in experimental systems. As with this work by Aakre et al., development of other new approaches may be essential for dissecting additional features in the evolution of protein-protein interactions.

REFERENCES

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Cell *163*, this issue, 594–606.

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., and Tawfik, D.S. (2005). Nat. Genet. *37*, 73–76.

Kim, J., and Copley, S.D. (2012). Proc. Natl. Acad. Sci. USA *109*, E2856–E2864.

Kortemme, T., Joachimiak, L.A., Bullock, A.N., Schuler, A.D., Stoddard, B.L., and Baker, D. (2004). Nat. Struct. Mol. Biol. *11*, 371–379.

Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., and Thornton, J.W. (2007). Science *317*, 1544–1548.

Smith, J.M. (1970). Nature *225*, 563–564.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Science *312*, 111–114.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Nature *426*, 676–680.

Zhu, L., Sharp, J.D., Kobayashi, H., Woychik, N.A., and Inouye, M. (2010). J. Biol. Chem. *285*, 39732–39738.

Bacterial Backstabbing: EF-Tu, Brute?

Matthew T. Cabeen[1] and Richard Losick[1,*]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

*Correspondence: losick@mcb.harvard.edu

http://dx.doi.org/10.1016/j.cell.2015.10.007

Bacterial type VI secretion is an offensive and defensive weapon that utilizes a molecular warhead to inject toxins into neighboring cells. In this issue of *Cell*, Whitney et al. report a new class of toxin that disrupts the core metabolism of recipient cells and uncover a surprising requirement for EF-Tu.

Nowhere is life more fiercely competitive than in the invisible world of bacteria and other microbes. Vast in numbers, these diminutive creatures disable their competitors by assailing each other with a range of weapons that include dispersible small molecules (antibiotics) and protein toxins (e.g., colicins). Perhaps the most cunning weapon is the type VI secretion (T6S) system of *Vibrio*, *Pseudomonas*, and certain other Gram-negative bacteria, which forms a miniscule spring-loaded dagger—a phage-tail-like contractile apparatus, complete with a molecularly poisoned sharp tip—to instantaneously inject protein toxins at point-blank range into neighboring cells. Evoking the infamous Umbrella Murder, in which Bulgarian dissident Georgi Markov was assassinated by a ricin-laced projectile fired from an umbrella, the spear-gun-like T6S system fires into eukaryotic cells and bacteria alike, breaching their membranes and delivering toxic effector molecules with different modes of action. In this issue of *Cell*, Whitney et al. (2015) report that a recently discovered effector poisons cells differently from previously known effectors and, surprisingly, requires the translation elongation factor EF-Tu to intoxicate target cells.

The first functionally characterized T6S system, from *V. cholerae*, was revealed by its role in warding off predation by amoebae (Pukatzki et al., 2006), but T6S systems are increasingly viewed as part of an arsenal that bacteria use against one another. Indeed, bacteria can even be seen duking it out, repeatedly attacking and counterattacking in a process termed "dueling" (Basler et al., 2013). Characterized T6S effector molecules include lipases that target the bacterial membrane, peptidoglycan hydrolases that degrade the cell wall, and nucleases that act on the nucleoid (Figure 1A) (Durand et al., 2014). Structural and mechanistic studies of a recently discovered effector, called Tse6, by Whitney et al. (2015) reveal yet a different mechanism. Tse6 resembles diphtheria toxin and other toxins that transfer ADP-ribose from NAD^+ onto proteins to inactivate them, but Tse6 is a pure glycohydrolase that intoxicates cells by depleting them of cytoplasmic nicotinamide adenine dinucleotide (phosphate) ($NAD(P)^+$). Attacker cells expressing Tse6 are protected by its cognate immunity protein, Tsi6, which tightly plugs the Tse6 active site.

An enduring question is where in the target cell the warhead of effector proteins is initially delivered. Is it to the periplasm only, to the cytoplasm, or to both? In principle, the phage-tail-like tube of the T6S apparatus is long enough to penetrate 500 nm into a target cell (Basler et al., 2012; Ho et al., 2014), which could allow for direct delivery into the cytoplasm. But lipases and

contributions of neutral drift and the impact of a few large-effect mutations versus many small-effect ones will need to be evaluated. Work on protein-protein interactions should be extended to other systems where cross-reaction is an issue, such as in other TA modules documented as cross-reacting (Zhu et al., 2010). Cross-reaction is also pertinent in other types of natural systems, as has been shown in some SH3 systems (Zarrinpar et al., 2003) and in the evolution of metabolic pathways (Kim and Copley, 2012). Ultimately, there are many other variables likely to be relevant in natural evolution that will surely be more difficult to ascertain in experimental systems. As with this work by Aakre et al., development of other new approaches may be essential for dissecting additional features in the evolution of protein-protein interactions.

REFERENCES

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Cell *163*, this issue, 594–606.

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., and Tawfik, D.S. (2005). Nat. Genet. *37*, 73–76.

Kim, J., and Copley, S.D. (2012). Proc. Natl. Acad. Sci. USA *109*, E2856–E2864.

Kortemme, T., Joachimiak, L.A., Bullock, A.N., Schuler, A.D., Stoddard, B.L., and Baker, D. (2004). Nat. Struct. Mol. Biol. *11*, 371–379.

Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., and Thornton, J.W. (2007). Science *317*, 1544–1548.

Smith, J.M. (1970). Nature *225*, 563–564.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Science *312*, 111–114.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Nature *426*, 676–680.

Zhu, L., Sharp, J.D., Kobayashi, H., Woychik, N.A., and Inouye, M. (2010). J. Biol. Chem. *285*, 39732–39738.

Bacterial Backstabbing: EF-Tu, Brute?

Matthew T. Cabeen[1] and Richard Losick[1,*]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

*Correspondence: losick@mcb.harvard.edu

http://dx.doi.org/10.1016/j.cell.2015.10.007

Bacterial type VI secretion is an offensive and defensive weapon that utilizes a molecular warhead to inject toxins into neighboring cells. In this issue of *Cell*, Whitney et al. report a new class of toxin that disrupts the core metabolism of recipient cells and uncover a surprising requirement for EF-Tu.

Nowhere is life more fiercely competitive than in the invisible world of bacteria and other microbes. Vast in numbers, these diminutive creatures disable their competitors by assailing each other with a range of weapons that include dispersible small molecules (antibiotics) and protein toxins (e.g., colicins). Perhaps the most cunning weapon is the type VI secretion (T6S) system of *Vibrio*, *Pseudomonas*, and certain other Gram-negative bacteria, which forms a miniscule spring-loaded dagger—a phage-tail-like contractile apparatus, complete with a molecularly poisoned sharp tip—to instantaneously inject protein toxins at point-blank range into neighboring cells. Evoking the infamous Umbrella Murder, in which Bulgarian dissident Georgi Markov was assassinated by a ricin-laced projectile fired from an umbrella, the spear-gun-like T6S system fires into eukaryotic cells and bacteria alike, breaching their membranes and delivering toxic effector molecules with different modes of action. In this issue of *Cell*, Whitney et al. (2015) report that a recently discovered effector poisons cells differently from previously known effectors and, surprisingly, requires the translation elongation factor EF-Tu to intoxicate target cells.

The first functionally characterized T6S system, from *V. cholerae*, was revealed by its role in warding off predation by amoebae (Pukatzki et al., 2006), but T6S systems are increasingly viewed as part of an arsenal that bacteria use against one another. Indeed, bacteria can even be seen duking it out, repeatedly attacking and counterattacking in a process termed "dueling" (Basler et al., 2013). Characterized T6S effector molecules include lipases that target the bacterial membrane, peptidoglycan hydrolases that degrade the cell wall, and nucleases that act on the nucleoid (Figure 1A) (Durand et al., 2014). Structural and mechanistic studies of a recently discovered effector, called Tse6, by Whitney et al. (2015) reveal yet a different mechanism. Tse6 resembles diphtheria toxin and other toxins that transfer ADP-ribose from NAD^+ onto proteins to inactivate them, but Tse6 is a pure glycohydrolase that intoxicates cells by depleting them of cytoplasmic nicotinamide adenine dinucleotide (phosphate) ($NAD(P)^+$). Attacker cells expressing Tse6 are protected by its cognate immunity protein, Tsi6, which tightly plugs the Tse6 active site.

An enduring question is where in the target cell the warhead of effector proteins is initially delivered. Is it to the periplasm only, to the cytoplasm, or to both? In principle, the phage-tail-like tube of the T6S apparatus is long enough to penetrate 500 nm into a target cell (Basler et al., 2012; Ho et al., 2014), which could allow for direct delivery into the cytoplasm. But lipases and

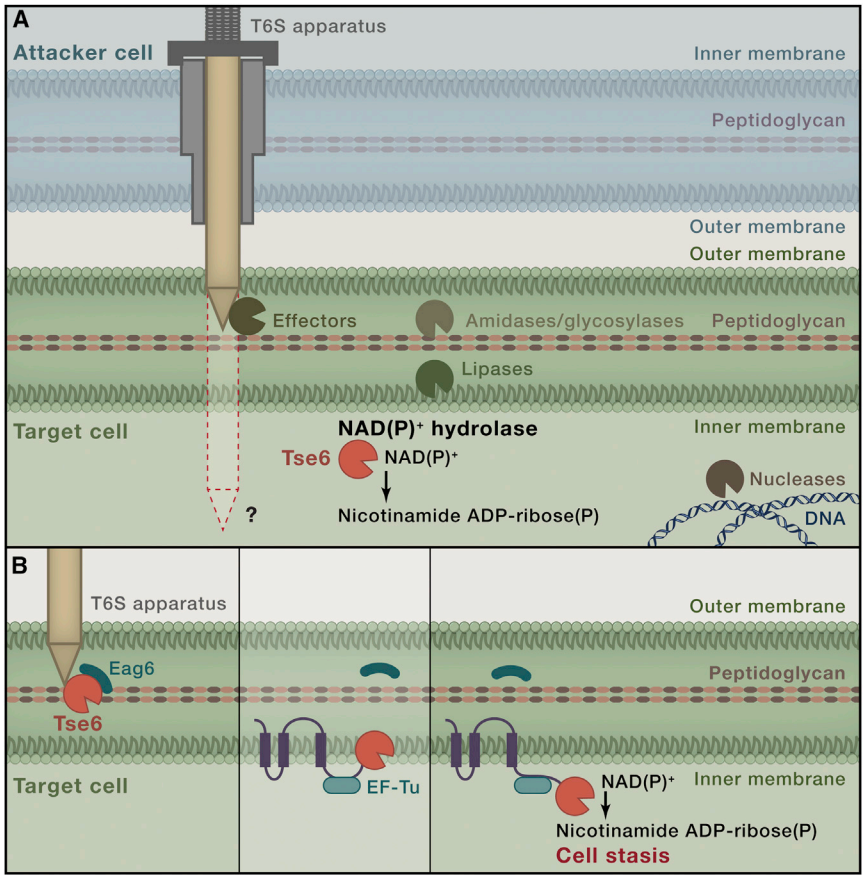


Figure 1. T6S Effectors and Their Sites of Action

(A) The T6S apparatus delivers effectors to the periplasmic space of Gram-negative target bacteria. It is unknown whether it can also breach the peptidoglycan and deliver effectors directly to the cytoplasm. Some effectors, such as lipases and peptidoglycan hydrolases (amidases or glycosylases), act within the periplasmic space. Others, such as nucleases and the newly discovered NAD(P)$^+$ hydrolase Tse6, act in the cytoplasm.

(B) Model for delivery and translocation of the Tse6 effector. When loaded on the T6S apparatus, the hydrophobic transmembrane regions of Tse6 are protected by its chaperone, Eag6 (left). Either in transit or once delivered, Eag6 dissociates from Tse6, exposing its three transmembrane regions. Tse6 then partitions into the cytoplasmic membrane, where its C-terminal effector domain is drawn into the cytoplasm by association with cytoplasmic EF-Tu (center). After being pulled into the cytoplasm, the effector intoxicates the target cell via its NAD(P)$^+$ hydrolase activity (right).

peptidoglycan hydrolases only need to reach the periplasmic space, and the cell wall might present a barrier to mechanical puncture. On the other hand, nuclease effectors and the Tse6 NAD(P)$^+$ hydrolase require entry into the cytoplasm to reach their targets. Until an attacking cell is visualized during the act of firing into its target, we will be forced to make inferences about how the toxins reach their targets. However, a surprising requirement of Tse6 sheds light on the mechanism of effector delivery for at least one toxin.

While investigating the interaction of Tse6 with other cellular proteins, the authors discovered that the NAD(P)$^+$ hydrolase forms a complex with the translation elongation factor EF-Tu. The interaction is strong and specific, and substitution of a single amino acid, identified in a Tse6-EF-Tu co-crystal structure, abolished Tse6 binding to EF-Tu. Remarkably, when this mutant Tse6 was introduced into attacker cells, they could no longer disable target cells, suggesting a specific requirement for EF-Tu in the delivery or activity of the toxin.

How are we to understand the unexpected requirement for EF-Tu in Tse6 toxicity? By using the non-interacting mutant, the authors ruled out several possibilities. Interaction with EF-Tu is required neither for the stability of the Tse6 protein nor for its NAD(P)$^+$ hydrolase activity. They also showed that it is not required for export of Tse6 from attacker cells. Thus, by a process of elimination, the authors conclude that the binding of Tse6 to EF-Tu facilitates the entry of Tse6 into the cytoplasm of target cells. In their model, Tse6 is delivered by the T6S system into the periplasm of a target cell along with its chaperone (Eag6), which shields its hydrophobic transmembrane regions. Tse6 then inserts into the cytoplasmic membrane of the target cell and is granted entry to the cytoplasm by progressively interacting with EF-Tu as it crosses the membrane (Figure 1B). Thus, the T6S system does not deliver the toxin into the cytoplasm. Rather, Tse6 only needs to reach the cytoplasmic membrane of the target cell, where it then becomes trapped in the cytoplasm by EF-Tu. If this model is correct, a complementary mutant of EF-Tu that is specifically blocked in binding to Tse6 would be expected to confer immunity to the toxin in target cells.

The Tse6 results do not exclude the possibility that other T6S effectors with cytoplasmic targets are directly delivered to the target cytoplasm, as Tse6 may have evolved a special, EF-Tu-dependent mode of cytoplasmic entry that is not shared by other cytoplasmic effectors. Perhaps other effectors, such as the T6S-delivered nuclease of *P. aeruginosa* (Hachani et al., 2014), are injected directly into the cytoplasm without the aid of target cell proteins. Finally, we note that T6S systems are not restricted to the delivery of toxins, as in the fascinating case of the transfer of proteins that mediate self/non-self identity between cells of *Proteus mirabilis* (Wenren et al., 2013).

REFERENCES

Basler, M., Pilhofer, M., Henderson, G.P., Jensen, G.J., and Mekalanos, J.J. (2012). Nature *483*, 182–186.

Basler, M., Ho, B.T., and Mekalanos, J.J. (2013). Cell *152*, 884–894.

Durand, E., Cambillau, C., Cascales, E., and Journet, L. (2014). Trends Microbiol. *22*, 498–507.

Hachani, A., Allsopp, L.P., Oduko, Y., and Filloux, A. (2014). J. Biol. Chem. *289*, 17872–17884.

Ho, B.T., Dong, T.G., and Mekalanos, J.J. (2014). Cell Host Microbe *15*, 9–21.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. (2006). Proc. Natl. Acad. Sci. USA *103*, 1528–1533.

Wenren, L.M., Sullivan, N.L., Cardarelli, L., Septer, A.N., and Gibbs, K.A. (2013). MBio *4*, e00374-13.

Whitney, J.C., Quentin, D., Sawai, S., LeRoux, M., Harding, B.N., Ledvina, H.E., Tran, B.Q., Robinson, H., Goo, Y.A., Goodlett, D.R., et al. (2015). Cell *163*, this issue, 607–619.

A Futile Approach to Fighting Obesity?

Shannon M. Reilly[1] and Alan R. Saltiel[1,*]
[1]Department of Medicine, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
*Correspondence: asaltiel@ucsd.edu
http://dx.doi.org/10.1016/j.cell.2015.10.006

The current obesity epidemic has focused a great deal of attention on mechanisms controlling energy balance. While diet and nutrient absorption affect energy intake, on the other side of the equation, energy expenditure is determined by basal metabolism, physical activity, and adaptive thermogenesis. Given various challenges in modulating these energy balance mechanisms to combat human obesity, many efforts have concentrated on how it might be possible to achieve weight loss through increased thermogenesis. In this issue of *Cell*, Kazak et al. describe a previously unrecognized molecular pathway for thermogenesis in fat cells.

Non-shivering thermogenesis occurs primarily in brown adipose tissue in rodents, but also has been detected in so called ''beige'' adipocytes, thought to reside mainly in subcutaneous fat tissue interspersed with classic white adipocytes (Wu et al., 2012; Young et al., 1984). Much of human thermogenic fat most closely resembles the rodent beige adipose tissue (Shinoda et al., 2015). Beige and brown adipocytes both express uncoupling protein 1 (Ucp1), which resides on the inner mitochondrial membrane. While the electron transport chain drives protons into the intermembrane space in mitochondria, creating a proton gradient across the inner membrane to drive the synthesis of ATP (Figure 1A), Ucp1 creates a pore through which protons disperse into the mitochondrial matrix, thereby generating heat and uncoupling ATP synthesis (Figure 1B). Cold exposure or increased sympathetic activity stimulated by feeding activates thermogenesis through adrenergic activation of *Ucp1* expression (Ricquier et al., 1986; Scarpace et al., 1997).

Although Ucp1 is well established as an important component of thermogenesis, investigators have long known that the transcriptional regulation of *Ucp1* cannot fully explain thermic responses. For example, the thermic effect of feeding is far too rapid to be explained by a transcriptional effect alone (Scarpace et al., 1997). Furthermore, *Ucp1* knockout mice can adapt to chronic cold exposure when the temperature transition is gradual (Golozoubova et al., 2001). Non-shivering thermogenesis has also been characterized in muscle, where Sarcolipin (*Sln*) uncouples ATP hydrolysis from Ca^{2+} transport, thereby creating a futile cycle that generates heat (Bal et al., 2012). However, *Ucp1/Sln* double-knockout mice still retain the ability to maintain thermal regulation when slowly adapted to the cold (Rowland et al., 2015), leaving a gap in our understanding of how thermogenesis occurs.

To fill in this gap, Bruce Spiegelman and his colleagues, including mass spec expert Steven Gygi, conducted proteomic and genomic studies comparing beige, white, and brown adipocytes (Kazak et al., 2015). KEGG pathway analysis of proteins preferentially expressed in beige versus brown fat revealed several components of the arginine/creatine and proline metabolism pathways. These findings were confirmed when the analysis was limited to proteins specifically enriched in purified mitochondrial fractions. Proteins that promote both creatine synthesis and phosphorylation, including the mitochondrial creatine kinase CMKT2 and the majority of ATP synthase subunits, were elevated in mitochondria from beige fat. Creatine kinase (CK) activity was also specifically induced in beige fat mitochondria derived from mice exposed to cold, suggesting that it is somehow under adrenergic control. Together, these findings hinted that a futile creatine phosphorylation and dephosphorylation cycle might somehow be involved in generating heat specifically in mitochondria from beige adipocytes.

CK catalyzes the phosphorylation of creatine using ATP, generating phosphocreatine and ADP. In tissues with high ATP demands, such as skeletal muscle, the high-energy phosphate bound to creatine can be transferred to ADP to generate cytosolic ATP (Wyss and Kaddurah-Daouk, 2000). If creatine were serving to regenerate mitochondrial ATP through classical CK-mediated phosphotransferase activity, it would be expected to boost respiration as a molar equivalent

Ho, B.T., Dong, T.G., and Mekalanos, J.J. (2014). Cell Host Microbe *15*, 9–21.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. (2006). Proc. Natl. Acad. Sci. USA *103*, 1528–1533.

Wenren, L.M., Sullivan, N.L., Cardarelli, L., Septer, A.N., and Gibbs, K.A. (2013). MBio *4*, e00374-13.

Whitney, J.C., Quentin, D., Sawai, S., LeRoux, M., Harding, B.N., Ledvina, H.E., Tran, B.Q., Robinson, H., Goo, Y.A., Goodlett, D.R., et al. (2015). Cell *163*, this issue, 607–619.

A Futile Approach to Fighting Obesity?

Shannon M. Reilly[1] and Alan R. Saltiel[1,*]
[1]Department of Medicine, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
*Correspondence: asaltiel@ucsd.edu
http://dx.doi.org/10.1016/j.cell.2015.10.006

The current obesity epidemic has focused a great deal of attention on mechanisms controlling energy balance. While diet and nutrient absorption affect energy intake, on the other side of the equation, energy expenditure is determined by basal metabolism, physical activity, and adaptive thermogenesis. Given various challenges in modulating these energy balance mechanisms to combat human obesity, many efforts have concentrated on how it might be possible to achieve weight loss through increased thermogenesis. In this issue of *Cell*, Kazak et al. describe a previously unrecognized molecular pathway for thermogenesis in fat cells.

Non-shivering thermogenesis occurs primarily in brown adipose tissue in rodents, but also has been detected in so called ''beige'' adipocytes, thought to reside mainly in subcutaneous fat tissue interspersed with classic white adipocytes (Wu et al., 2012; Young et al., 1984). Much of human thermogenic fat most closely resembles the rodent beige adipose tissue (Shinoda et al., 2015). Beige and brown adipocytes both express uncoupling protein 1 (Ucp1), which resides on the inner mitochondrial membrane. While the electron transport chain drives protons into the intermembrane space in mitochondria, creating a proton gradient across the inner membrane to drive the synthesis of ATP (Figure 1A), Ucp1 creates a pore through which protons disperse into the mitochondrial matrix, thereby generating heat and uncoupling ATP synthesis (Figure 1B). Cold exposure or increased sympathetic activity stimulated by feeding activates thermogenesis through adrenergic activation of *Ucp1* expression (Ricquier et al., 1986; Scarpace et al., 1997).

Although Ucp1 is well established as an important component of thermogenesis, investigators have long known that the transcriptional regulation of *Ucp1* cannot fully explain thermic responses. For example, the thermic effect of feeding is far too rapid to be explained by a transcriptional effect alone (Scarpace et al., 1997). Furthermore, *Ucp1* knockout mice can adapt to chronic cold exposure when the temperature transition is gradual (Golozoubova et al., 2001). Non-shivering thermogenesis has also been characterized in muscle, where Sarcolipin (*Sln*) uncouples ATP hydrolysis from Ca^{2+} transport, thereby creating a futile cycle that generates heat (Bal et al., 2012). However, *Ucp1/Sln* double-knockout mice still retain the ability to maintain thermal regulation when slowly adapted to the cold (Rowland et al., 2015), leaving a gap in our understanding of how thermogenesis occurs.

To fill in this gap, Bruce Spiegelman and his colleagues, including mass spec expert Steven Gygi, conducted proteomic and genomic studies comparing beige, white, and brown adipocytes (Kazak et al., 2015). KEGG pathway analysis of proteins preferentially expressed in beige versus brown fat revealed several components of the arginine/creatine and proline metabolism pathways. These findings were confirmed when the analysis was limited to proteins specifically enriched in purified mitochondrial fractions. Proteins that promote both creatine synthesis and phosphorylation, including the mitochondrial creatine kinase CMKT2 and the majority of ATP synthase subunits, were elevated in mitochondria from beige fat. Creatine kinase (CK) activity was also specifically induced in beige fat mitochondria derived from mice exposed to cold, suggesting that it is somehow under adrenergic control. Together, these findings hinted that a futile creatine phosphorylation and dephosphorylation cycle might somehow be involved in generating heat specifically in mitochondria from beige adipocytes.

CK catalyzes the phosphorylation of creatine using ATP, generating phosphocreatine and ADP. In tissues with high ATP demands, such as skeletal muscle, the high-energy phosphate bound to creatine can be transferred to ADP to generate cytosolic ATP (Wyss and Kaddurah-Daouk, 2000). If creatine were serving to regenerate mitochondrial ATP through classical CK-mediated phosphotransferase activity, it would be expected to boost respiration as a molar equivalent

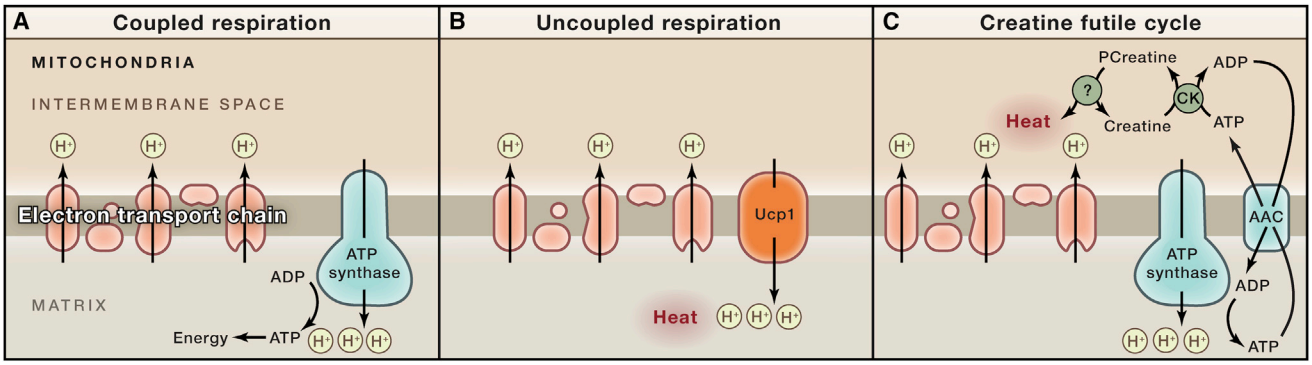


Figure 1. Different Modes of Mitochondrial Respiration

(A) Coupled respiration, which generates ATP.

(B) Thermogenesis through uncoupled respiration by Ucp1, which does not involve ATP synthase.

(C) Thermogenesis by creatine futile cycle, which requires ATP synthase activity, although no net ATP is generated.

to ADP, when ADP concentrations are limiting. Consistent with this classical function of creatine, addition of 0.01 mM creatine in the presence of 1 mM ADP to mitochondria isolated from classic brown fat and muscle had no detectable effect on respiration. However, in beige fat mitochondria, this small amount of creatine produced a large effect on respiratory rate, far exceeding that expected from 1:1 stoichiometry with ADP, suggesting that creatine is regenerated from phosphocreatine via a futile cycle that dissipates the energy as heat (Figure 1C). This idea was supported by direct calorimetry, which demonstrated that addition of small amounts of creatine increased heat production in beige, but not brown, mitochondria. In contrast to Ucp1-mediated thermogenesis, this futile creatine cycle requires coupled ATP synthesis, although no net ATP is generated.

The identification of this futile cycle may advance our understanding of beige-fat-specific thermogenesis in adult humans who possess little if any BAT. However, numerous questions remain about the molecular mechanics of this futile creatine cycle. Notable among these is the mechanism of dephosphorylation. While Kazak et al. note that the mitochondrial phosphatase Phospho1 exhibits an expression pattern that suggests its participation in this cycle, this phosphatase did not catalyze dephosphorylation of phospho-creatine in vitro. The authors propose that Phospho1 may play a unique role at the end of a phosphotransfer chain, but other players are probably involved, and it will be important to identify the relevant phosphatase or transferase(s) that complete this futile cycle.

An equally important question is how the transport and flux of creatine in mitochondria affects the activity of this futile cycle. In principle, even diminishingly small quantities of creatine could continually undergo phosphorylation and dephosphorylation, obviating the need for significant creatine synthesis in beige adipose tissue. Indeed, creatine levels are an order of magnitude higher in brown fat, where this futile cycle does not appear to be active. Along the same lines, it is not clear why the creatine transport inhibitor, β-GPA, which reduces creatine levels by less than 50%, would have such a profound effect on beige fat thermogenesis, as it reduced oxygen consumption in response to β-adrenergic stimulation in beige fat, as well as core body temperature of cold-adapted *Ucp1* knockout mice. Finally, a key question concerns how the cycle may be regulated, particularly in response to adrenergic activation of the beige fat cell.

Additional investigations into this futile cycle by genetic and pharmacological manipulation of its activity will hopefully reveal its relative contribution to energy expenditure in humans and whether or not it is modulated in obesity. If it does prove to be an important component of adaptive thermogenesis, therapeutic or even dietary agents might be employed to activate the process and perhaps achieve weight loss in obese individuals. Let's hope that efforts to decipher the mechanisms of this cycle are not futile.

REFERENCES

Bal, N.C., Maurya, S.K., Sopariwala, D.H., Sahoo, S.K., Gupta, S.C., Shaikh, S.A., Pant, M., Rowland, L.A., Bombardier, E., Goonasekera, S.A., et al. (2012). Nat. Med. *18*, 1575–1579.

Golozoubova, V., Hohtola, E., Matthias, A., Jacobsson, A., Cannon, B., and Nedergaard, J. (2001). FASEB J. *15*, 2048–2050.

Kazak, L., Chouchani, E.T., Jedrychowski, M.P., Erickson, B.K., Shinoda, K., Cohen, P., Vetrivela, R., Lu, G.Z., Laznik-Bogoslavski, D., Hasenfuss, S., et al. (2015). Cell *163*, this issue, 643–655.

Ricquier, D., Bouillaud, F., Toumelin, P., Mory, G., Bazin, R., Arch, J., and Pénicaud, L. (1986). J. Biol. Chem. *261*, 13905–13910.

Rowland, L.A., Bal, N.C., Kozak, L.P., and Periasamy, M. (2015). J. Biol. Chem. *290*, 12282–12289.

Scarpace, P.J., Matheny, M., Pollock, B.H., and Tümer, N. (1997). Am. J. Physiol. *273*, E226–E230.

Shinoda, K., Luijten, I.H., Hasegawa, Y., Hong, H., Sonne, S.B., Kim, M., Xue, R., Chondronikola, M., Cypess, A.M., Tseng, Y.H., et al. (2015). Nat. Med. *21*, 389–394.

Wu, J., Boström, P., Sparks, L.M., Ye, L., Choi, J.H., Giang, A.H., Khandekar, M., Virtanen, K.A., Nuutila, P., Schaart, G., et al. (2012). Cell *150*, 366–376.

Wyss, M., and Kaddurah-Daouk, R. (2000). Physiol. Rev. *80*, 1107–1213.

Young, P., Arch, J.R., and Ashwell, M. (1984). FEBS Lett. *167*, 10–14.

Cell

Imaging the Neural Basis of Locomotion

Kristin Branson[1,*] and Jeremy Freeman[1]
[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA
*Correspondence: bransonk@janelia.hhmi.org
http://dx.doi.org/10.1016/j.cell.2015.10.014

To investigate the fundamental question of how nervous systems encode, organize, and sequence behaviors, Kato et al. imaged neural activity with cellular resolution across the brain of the worm *Caenorhabditis elegans*. Locomotion behavior seems to be continuously represented by cyclical patterns of distributed neural activity that are present even in immobilized animals.

The worm *Caenorhabditis elegans* might seem simple, but there's a lot going on both inside and out. It explores the world with a series of sinusoidal swims punctuated by pirouettes—quick reversals or sharp turns. Poking the animal reliably evokes a reflexive reversal (de Bono and Maricq, 2005). Other stimuli, like chemicals, can elicit more stochastic changes, for example, when they suggest the presence of food. When worms detect decreases in the amount of food nearby, the probability they will pirouette increases (de Bono and Maricq, 2005). This behavior program helps worms navigate toward areas with high food concentration in a non-deterministic manner that's difficult for predators to exploit. How does *C. elegans* produce variable behavior sequences? How does it integrate diverse sensory information to adjust the probabilities of its behaviors? Understanding how a nervous system encodes, organizes, and sequences behaviors is a central problem in systems neuroscience—whether in worms or humans. In this issue of *Cell*, Kato et al. (2015) shed new light on how neural dynamics in the worm produce behavior. By combining several new technologies, including near-whole-brain imaging of neural activity at single-cell resolution, Kato et al. (2015) provide evidence that motor commands in *C. elegans* are represented across large populations of neurons with cyclical dynamics, building on and extending related observations across a variety of behaviors such as digestion (Marder and Calabrese, 1996) and decision-making and across organisms ranging from leeches (Briggman et al., 2005) to primates (Churchland et al., 2012).

Kato et al. (2015) imaged neural activity across ~100 neurons in the brains of worms held immobile in a small channel. They used GCAMP, a genetically encoded calcium indicator that elicits green fluorescence in active neurons, and a fast, commercially available, confocal microscope to capture volumetric images of the entire brain every third of a second. More sophisticated microscopes have been used to image neural activity in larger volumes at higher speeds (Prevedel et al., 2014), but Kato et al. (2015)'s system achieved single-neuron resolution across most of the brain and, when combined with the extensive existing knowledge of *C. elegans* neural anatomy, supported identification of most neurons.

Despite being held immobile, the worms' brains fluttered with activity over long durations (~20 min per worm). The resulting data—100-dimensional time series describing neural activity, one dimension for each cell—are difficult to understand with the naked eye and require higher-level analysis. Kato et al. (2015) employed principal component analysis (PCA) to reduce the high-dimensional data to two- or three-dimensional trajectories (Figure 1A). These dynamical portraits captured the structure of a neuronal population in a simpler and more interpretable form, revealing trajectories of neuronal activity that followed a cyclical, highly repeatable pattern. That is, one stereotyped pattern of neural activity is followed by a second stereotyped pattern and so on, until the original pattern occurs again and the cycle repeats. In addition, Kato et al. (2015) observed that the activity of many neurons contributed to this neural representation, suggesting that these neural dynamics, though restricted to just a couple of dimensions, were distributed across a large number of neurons. Cyclical neural dynamics have been observed in the generation of rhythmic motor behaviors like digestion and swimming (Marder and Calabrese, 1996), as well as non-periodic behaviors like reaching for a target (Churchland et al., 2012). Low-dimensional, distributed neural representations also have been observed in many systems, including locomotion behavior choice (Briggman et al., 2005) and odor-identity encoding (Stopfer et al., 2003). Both oscillatory and distributed neural representations are hypothesized to be fundamental neural organizational strategies (Briggman and Kristan, 2008), about which numerous open questions remain. How does neural activity state relate to behavior? How are cyclical patterns generated and how do the many neurons in this network coordinate their activity? What are the advantages of these implementations? How are sensory information, learning, and the animal's history integrated to change the neural representation?

Through a combination of analysis and experiment—many exploiting the unique experimental capabilities in *C. elegans*—Kato et al. (2015) tried to demystify these neural trajectories with more concrete observations. They performed a second set of experiments in which they imaged neural activity in freely behaving worms to observe locomotion concomitant with neural activity. In these more limited recordings, they used genetic techniques to target calcium indicator expression to cells identified as important by the PCA analysis of whole-brain data. Surprisingly, clusters of the neural trajectory space, defined solely on the basis of neural activity, corresponded to different locomotion behaviors: swimming forward, reversing,

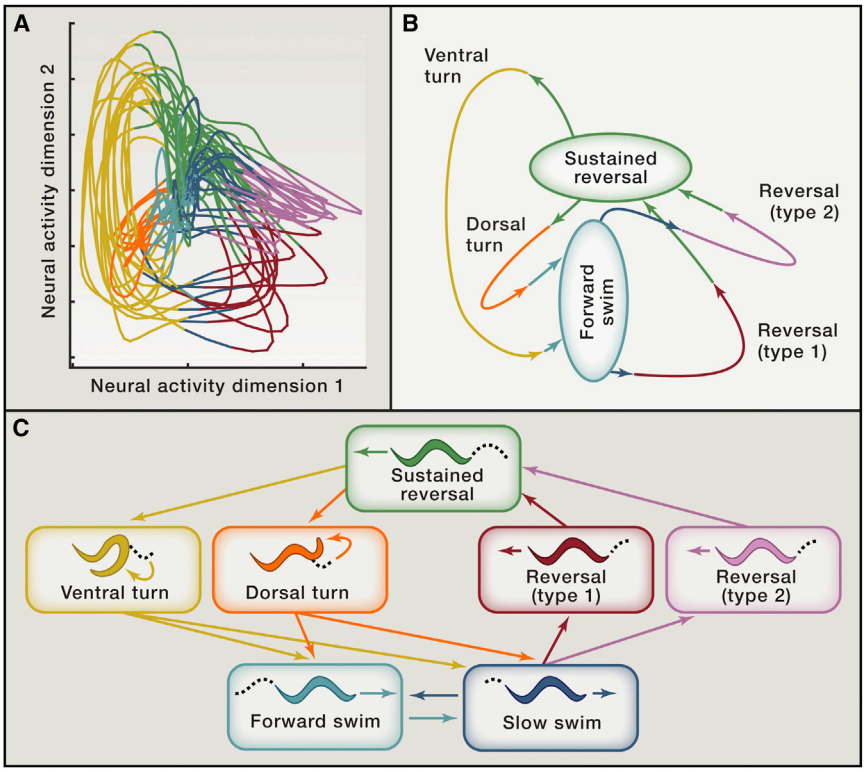


Figure 1. Neural Representation of Locomotion Behavior

(A) The trajectory of neural activity obtained by near-whole-brain-imaging at single-cell resolution in *C. elegans* projected onto the first two principal components. The dynamics are cyclical and stereotyped. Color indicates which cluster each segment of neural activity belongs to. Adapted from Kato et al. (2015).

(B) Regions of neural activity space annotated with the behaviors they correspond to. Adapted from Kato et al. (2015).

(C) State transition diagram describing *C. elegans* locomotion behavior paralleled by the neural dynamics shown in (B).

and ventral and dorsal turns (Figure 1B). Thus, it appears that a large portion of the *C. elegans* neural activity, even in immobilized animals, encodes the (fictive) locomotor state. The neural activity flow (Figure 1B) has similarities to a continuous version of a behavior state transition diagram (Figure 1C), a representation often used to visualize the types and probabilities of behavior transitions observed (Anderson and Perona, 2014).

A surprising feature of these patterns of neural activity is that they are largely self-generated. Eliminating activity in an output motor command neuron left much of the cyclical activity patterns intact, and environmental input also seemed to have limited influence on the shape of the neural activity manifold. Instead, increasing oxygen concentration increased the frequency with which activity entered regions of neural space associated with reversals. Taken together, these observations suggest that a large fraction of the brain of the worm is constantly oscillating between states which, when the animal is freely behaving, cause it to perform different locomotion behaviors. In this framework, sensory information modulates the probabilities of entering and leaving these states. Thus, these neural dynamics may explain the variability exhibited in the worm's locomotion behavior (Gordus et al., 2015).

Kato et al. (2015) have opened up whole-brain imaging in *C. elegans* as a new, powerful system for investigating how the brain sequences behavior, as well as how cyclical and distributed neural representations are generated and maintained. Although behaviorally simpler than animals like flies, mice, or primates, *C. elegans* has experimental advantages, many demonstrated here: transparency, developmental stereotypy, thoroughly characterized anatomy, and genetic control. In future research, this system might be enhanced if whole-brain imaging were possible in freely behaving animals, making the neural-behavioral relationships explored here more direct. It would also be aided by development of faster indicators of neural activity, other fluorescent sensors, and microscopes with subcellular resolution. New analytical methods for relating neural trajectories to behavior might also result in new discoveries and will become increasingly necessary as these techniques are extended to more complex animals like zebrafish, fruit flies, mice, and, eventually, primates and humans.

REFERENCES

Anderson, D.J., and Perona, P. (2014). Neuron *84*, 18–31.

Briggman, K.L., and Kristan, W.B., Jr. (2008). Annu. Rev. Neurosci. *31*, 271–294.

Briggman, K.L., Abarbanel, H.D., and Kristan, W.B., Jr. (2005). Science *307*, 896–901.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Nature *487*, 51–56.

de Bono, M., and Maricq, A.V. (2005). Annu. Rev. Neurosci. *28*, 451–501.

Gordus, A., Pokala, N., Levy, S., Flavell, S.W., and Bargmann, C.I. (2015). Cell *161*, 215–227.

Kato, S., Kaplan, H.S., Schrödel, T., Skora, S., Lindsay, T.H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Cell *163*, this issue, 656–669.

Marder, E., and Calabrese, R.L. (1996). Physiol. Rev. *76*, 687–717.

Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E.S., and Vaziri, A. (2014). Nat. Methods *11*, 727–730.

Stopfer, M., Jayaraman, V., and Laurent, G. (2003). Neuron *39*, 991–1004.

Cell

Ethylene Prunes Translation

Mohammad Salehin[1] and Mark Estelle[1,*]
[1]Howard Hughes Medical Institute and Section of Cell and Developmental Biology, University of California, San Diego, La Jolla, CA 92093, USA
*Correspondence: mestelle@ucsd.edu
http://dx.doi.org/10.1016/j.cell.2015.10.032

Ethylene regulates many aspects of plant growth and development. In the presence of ethylene, the C terminus of EIN2 (EIN2C) translocates into the nucleus and activates transcription. Li et al. and Merchante et al. show that EIN2C also regulates translation through an interaction with the 3′ UTRs of transcripts.

Ethylene is a gaseous plant hormone known to affect diverse aspects of plant growth and development, including leaf abscission, germination, leaf epinasty, senescence, and fruit ripening, as well as biotic and abiotic stress responses. The ethylene-signaling pathway in plants is well understood. Ethylene is perceived by the ETR1/ETR2/ERS1/ERS2/EIN4 receptors on the endoplasmic reticulum (ER) membrane. In the absence of ethylene, the receptors activate CTR1, a Ser/Thr kinase that suppresses the ethylene response. This is accomplished by phosphorylation of another ER membrane protein EIN2, a critical component of the ethylene signaling pathway (Alonso et al., 1999) (Figure 1A). In the presence of ethylene, EIN2 is no longer phosphorylated, and its C terminus (EIN2C) is cleaved by unknown proteases and translocated into the nucleus where it activates the master transcriptional regulators EIN3 and EIL1 (Qiao et al., 2012) (Figure 1B).

Not only the activity of EIN3 is modulated by ethylene signaling via the "cleave and shuttle" of EIN2—its abundance is also subjected to regulation by ethylene. EIN3 is a short-lived protein that is degraded by the ubiquitin proteasome system in the absence of ethylene, a process that is mediated by two ubiquitin E3 ligases containing the F-box proteins EBF1 and EBF2 (An et al., 2010) (Figure 1A). In the presence of ethylene, EIN2C promotes the degradation of EBF1/2 and EIN3 accumulates in the nucleus (Figure 1B).

Although EIN3/EIL1-dependent transcriptional regulation constitutes a major fraction of the ethylene response, the discovery that some rapid ethylene growth responses are EIN2 dependent but don't require EIN3/EIL1 led to speculation that the pathway branched after EIN2 (Binder et al., 2004). In this issue of *Cell*, two papers by Li et al. (2015) and Merchante et al. (2015) may have provided the molecular basis for this second pathway in *Arabidopsis*.

Merchante et al. (2015) start their study by asking if ethylene has any effect on the translation of specific genes. They use genome-wide ribosomal footprinting and RNA-seq to identify genes that are translationally regulated by ethylene. Interestingly, *EBF1/2* are among the genes, and their translation is downregulated by ethylene (Merchante et al., 2015). Li et al. (2015) come to the same conclusion by a different approach. They follow up on an earlier observation that mRNA fragments containing the 3′ UTR of *EBF1/2* accumulate in an ethylene-insensitive mutant and find that overexpression of this 3′ UTR results in the unresponsiveness to ethylene stimulation (Olmedo et al., 2006). This effect is due to increased translation of the *EBF1/2* mRNA in the presence of excess 3′ UTR and a subsequent decrease in EIN3 levels, suggesting that the 3′ UTR of *EBF1/2* is involved in ethylene-mediated translational control of EBF1/2.

Protein translational control has several advantages. The response to a signal can be very rapid, and because the mRNA template is not destroyed, regulation is easily reversible. The process is often mediated by the binding of regulatory proteins to the 5′ or 3′ UTR. Moreover, microRNAs can also bind to the 3′ UTR of the target RNA to control either mRNA decay or translation of the target protein (Jia et al., 2013, Mayr and Bartel, 2009). Often, mRNAs that are not translated aggregate into cytoplasmic mRNP granules, known as P-bodies and stress granules, where they may be degraded.

In the case of *EBF1/2*, both Li et al. (2015) and Merchante et al. (2015) show that the decreased translation of EBF1/2 by ethylene signaling is dependent on EIN2 but independent of EIN3/EIL1. Following up on this genetic evidence, the authors further demonstrate that EIN2C binds to the *EBF1/2* 3′ UTRs, either directly or indirectly, to regulate translation of the respective proteins (Figure 1B). Finally, the authors show that EIN2C, the *EBF1* 3′ UTR, and EIN5 all localize in cytoplasmic P-bodies upon ethylene stimulation (Figure 1B). Thus, it is possible that inhibition of translation is due to recruitment of the RNAs to the P-bodies.

In the paper by Merchante et al. (2015), these studies of translational regulation converge with a second project, a genetic screen to identify new genes that are required for ethylene response. Among the genes recovered in this screen are three members of the nonsense-mediated RNA decay machinery, *UPF1*, *UPF2*, and *UPF3*. Since the UPFs are known to inhibit translation, the effects of the *upf* mutants on *EBF1/2* translation were tested (Merchante et al., 2015). Indeed, the *upf2-10* mutation clearly reduces ethylene-dependent translational regulation of *EBF1/2* mRNAs. The authors also provide evidence that UPF2 co-localizes with EIN2C in the P-bodies and propose that the UPF proteins may facilitate the interaction between EIN2C and the 3′UTR of *EBF1/2* mRNAs.

These studies shed light on the function of the enigmatic EIN2 protein and deepen our understanding of the ethylene-signaling pathway. There are a number of

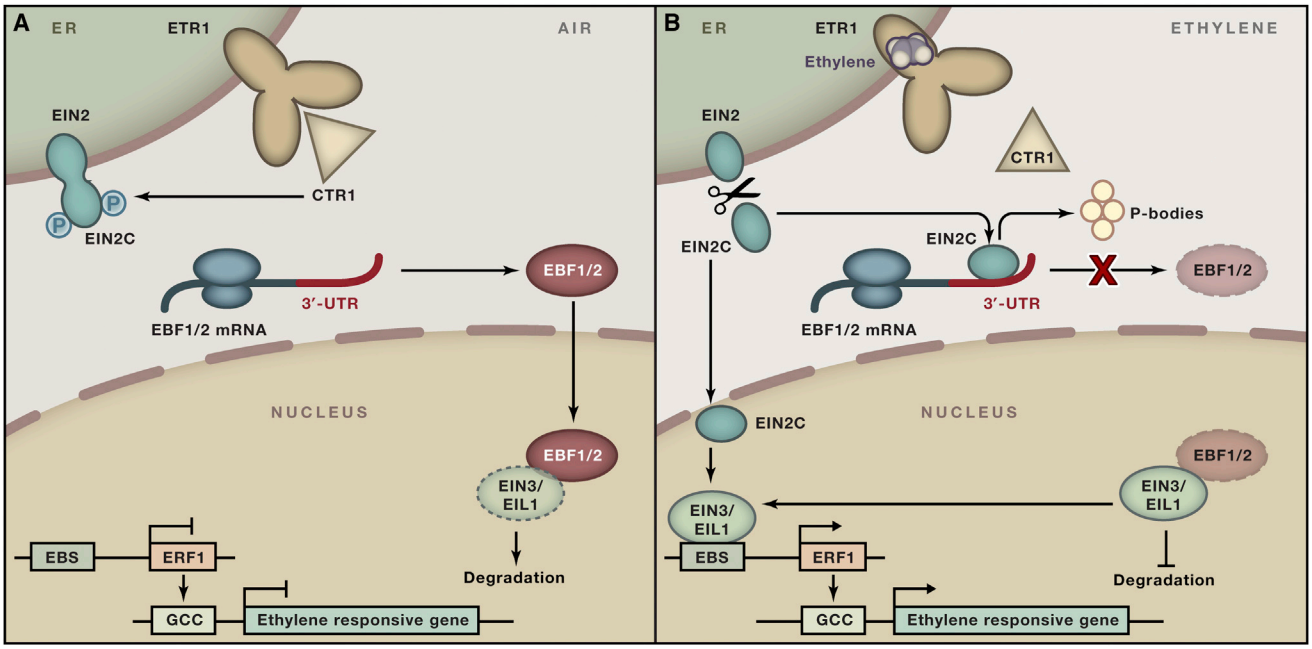


Figure 1. Transcriptional and Translational Regulation in the Ethylene Signaling Pathway

(A) In the absence of ethylene, ethylene receptors (ETR1/ETR2/ERS1/ERS2/EIN4) on the ER membrane activate CTR1, a Ser/Thr kinase, which phosphorylates another ER membrane protein EIN2C. The F-Box proteins EBF1 and EBF2 bind and degrade the master transcription factors EIN3 and EIL1 via the ubiquitin proteasome system, preventing the ethylene-stimulated transcription cascade.

(B) In the presence of ethylene, receptors perceive ethylene at the ER membrane and no longer activate CTR1. As a result, unphosphorylated EIN2 is cleaved by an unknown mechanism, and the cleavage product, the C terminus of EIN2 (EIN2C), shuttles to the nucleus, where it activates the master transcription factors EIN3/EIL1 and the downstream transcription cascade. Concurrently, cytoplasmic EIN2C directly or indirectly binds to the 3′ UTR of *EBF1* and *EBF2* mRNAs, which inhibits their translation.

interesting questions that remain to be answered. The precise mechanism of how EBF1/2 translational regulation is achieved by EIN2C is yet to be described. How EIN2C specifically recognizes the 3′ UTR of *EBF1/2* transcripts, what features of the UTR are important, and if and how additional factors are involved in the binding and subsequent translational repression warrant further investigation. It is possible that recruitment of the *EBF1/2* to the P-body is itself sufficient to inhibit translation (Maldonado-Bonilla, 2014). Alternatively, EIN2C, or another interacting protein such as UPF2 may directly inhibit translation, perhaps by preventing the formation of the 43S initiation complex. Another important question is that of the fate of *EBF1/2* RNAs once they reach the P-body. Are they degraded or can they be released and subsequently translated anew? Finally, it is still not clear what factors downstream of EIN2 mediate the rapid ethylene response. Although the effect of ethylene on EBF1/2 translation reported in these studies is independent of EIN3/EIL1, because EBF1/2 regulate the EIN3 protein level, the outcome of EBF1/2 translational regulation still requires EIN3. Presumably, other targets of EIN2-dependent translational regulation are responsible for the rapid ethylene response observed previously (Binder et al., 2004). In any case, an enhanced understanding of ethylene signaling may have important practical benefits. Many aspects of plant physiology and development are mediated by ethylene, and the ability to manipulate the pathway in crops is likely to lead to important improvements in crop yield.

REFERENCES

Alonso, J.M., Hirayama, T., Roman, G., Nourizadeh, S., and Ecker, J.R. (1999). Science *284*, 2148–2152.

An, F., Zhao, Q., Ji, Y., Li, W., Jiang, Z., Yu, X., Zhang, C., Han, Y., He, W., Liu, Y., et al. (2010). Plant Cell *22*, 2384–2401.

Binder, B.M., Mortimore, L.A., Stepanova, A.N., Ecker, J.R., and Bleecker, A.B. (2004). Plant Physiol. *136*, 2921–2927.

Jia, J., Yao, P., Arif, A., and Fox, P.L. (2013). Curr. Opin. Genet. Dev. *23*, 29–34.

Li, W., Ma, M., Feng, Y., Li, H., Wang, Y., Ma, Y., Li, M., An, F., and Guo, H. (2015). Cell *163*, this issue, 670–683.

Maldonado-Bonilla, L.D. (2014). Front. Plant Sci. *5*, 201.

Mayr, C., and Bartel, D.P. (2009). Cell *138*, 673–684.

Merchante, C., Brumos, J., Yun, J., Hu, Q., Spencer, K.R., Enriquez, P., Binder, B.M., Heber, S., Stepanova, A.N., and Alonso, J.M. (2015). Cell *163*, this issue, 684–697.

Olmedo, G., Guo, H., Gregory, B.D., Nourizadeh, S.D., Aguilar-Henonin, L., Li, H., An, F., Guzman, P., and Ecker, J.R. (2006). Proc. Natl. Acad. Sci. USA *103*, 13286–13293.

Qiao, H., Shen, Z., Huang, S.S., Schmitz, R.J., Urich, M.A., Briggs, S.P., and Ecker, J.R. (2012). Science *338*, 390–393.

Leading Edge
Minireview

Cell

Germinal Center Selection and the Antibody Response to Influenza

Gabriel D. Victora[1,*] and Patrick C. Wilson[2,*]
[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Department of Medicine, Section of Rheumatology, The Committee on Immunology, The Knapp Center for Lupus and Immunology Research, The University of Chicago, Chicago, IL 60637, USA
*Correspondence: victora@wi.mit.edu (G.D.V.), wilsonp@uchicago.edu (P.C.W.)
http://dx.doi.org/10.1016/j.cell.2015.10.004

In this Minireview, we discuss basic aspects of germinal center biology in the context of immunity to influenza infection and speculate on how the simultaneous evolutionary races of virus and antibody may impact our efforts to design a universal influenza vaccine.

Introduction

Influenza epidemics cause millions of infections and hundreds of thousands of deaths worldwide each year and cost nearly $100 billion per year in the United States alone. The influenza vaccine is generally protective against the strains from which it is composed. However, effectiveness wanes as herd immunity pushes the viral envelope proteins to mutate and evolve (antigenic drift). Periodically, more antigenically distinct or virulent influenza strains arise due to recombination among zoonotic strains (antigenic shift). These strains can cause pandemics such as the 1918 Spanish flu, which had a death toll of tens of millions of people.

The primary target of anti-influenza antibodies is the hemagglutinin (HA) protein, a trimer consisting of a membrane (envelope)-embedded stalk region and an expanded globular head on which the receptor-binding site (RBS) is located. Most protective antibodies against HA bind to regions surrounding the RBS that are highly mutable, which allows antigenic drift and immune escape. However, rare antibodies have been isolated that bind functionally critical regions of HA that are much less susceptible to antigenic drift (Krammer and Palese, 2015; Schmidt et al., 2015). These antibodies bind either within the RBS, mimicking the sialic acid ligands of HA, or to regions of the HA stalk that are critical for viral fusion to host cell membranes (Figure 1A). A major goal of vaccinologists is to develop a universal vaccine capable of eliciting protective antibodies to epitopes that are common among influenza strains and that are stable over time, thus circumventing antigenic variation (Krammer and Palese, 2015).

Antibodies attain high affinity through somatic hypermutation (SHM) of immunoglobulin (Ig) genes in B cells following exposure to antigen in a process known as affinity maturation (Eisen, 2014). Most antibodies to influenza cloned from humans are heavily mutated, and these mutations are likely critical for broadly protective binding to the virus (Lingwood et al., 2012; Pappas et al., 2014; Schmidt et al., 2015). Affinity maturation takes place in germinal centers (GCs) (Victora and Nussenzweig, 2012), where B cells undergo SHM and are subsequently selected based on the ability of their mutant Igs to bind antigen. A fundamental constraint to this process is that GCs select for antibodies with higher affinity for antigen (or some close correlate of it) but are "agnostic" when it comes to their protective efficacy—including their ability to neutralize virus or kill infected cells and their potential to cross-react with other strains of the offending pathogen (breadth). For many infectious diseases, this "evolution by proxy" is sufficient to provide robust immunity. However, in cases like influenza, antigenic drift renders high-affinity protective antibodies from one season ineffective against newly emerging strains.

GC Kinetics and Structure

Antigenic stimulation triggers specific B and T cells to move toward the T zone/follicle (T:B) border area of secondary lymphoid organs. There, B cells that present antigen-derived peptides to helper T cells become "authorized" to engage in a productive immune response. Successful B cells enter one of three developmental paths: they can differentiate into plasma cells (PCs) that secrete early, low-affinity antibody; they can re-establish a nonproliferative state and join the memory B cell pool; or they can enter the GC reaction (Figure 1B) (Victora and Nussenzweig, 2012).

GCs appear several days after antigen exposure as clusters of rapidly proliferating cells in the center of B cell follicles. GCs comprise two anatomically defined areas: the dark zone (DZ), where cells proliferate and hypermutate their *Ig* genes, and the light zone (LZ), where antigen-driven selection takes place (Victora and Nussenzweig, 2012). Following DZ hypermutation, B cells migrate to the LZ, where antigen is deposited as immune complexes on the surface of follicular dendritic cells (FDCs). LZ B cells compete to bind and retrieve antigen from FDCs and present it to GC-resident T follicular helper (Tfh) cells. B cells that have acquired higher affinity by virtue of SHM are more likely to receive positive selection signals, triggering their return to the DZ for further proliferation and hypermutation (cyclic re-entry, Figure 1B). GC selection is thus reminiscent of Darwinian evolution: iterative cycles of descent with modification (SHM) followed by fitness (affinity)-based selection lead to increased fitness of the population as a whole. Sporadic differentiation of positively selected LZ B cells into PCs and memory B cells results in the progressive increase in the affinity of serum antibodies over time and upon re-immunization (Figure 1B).

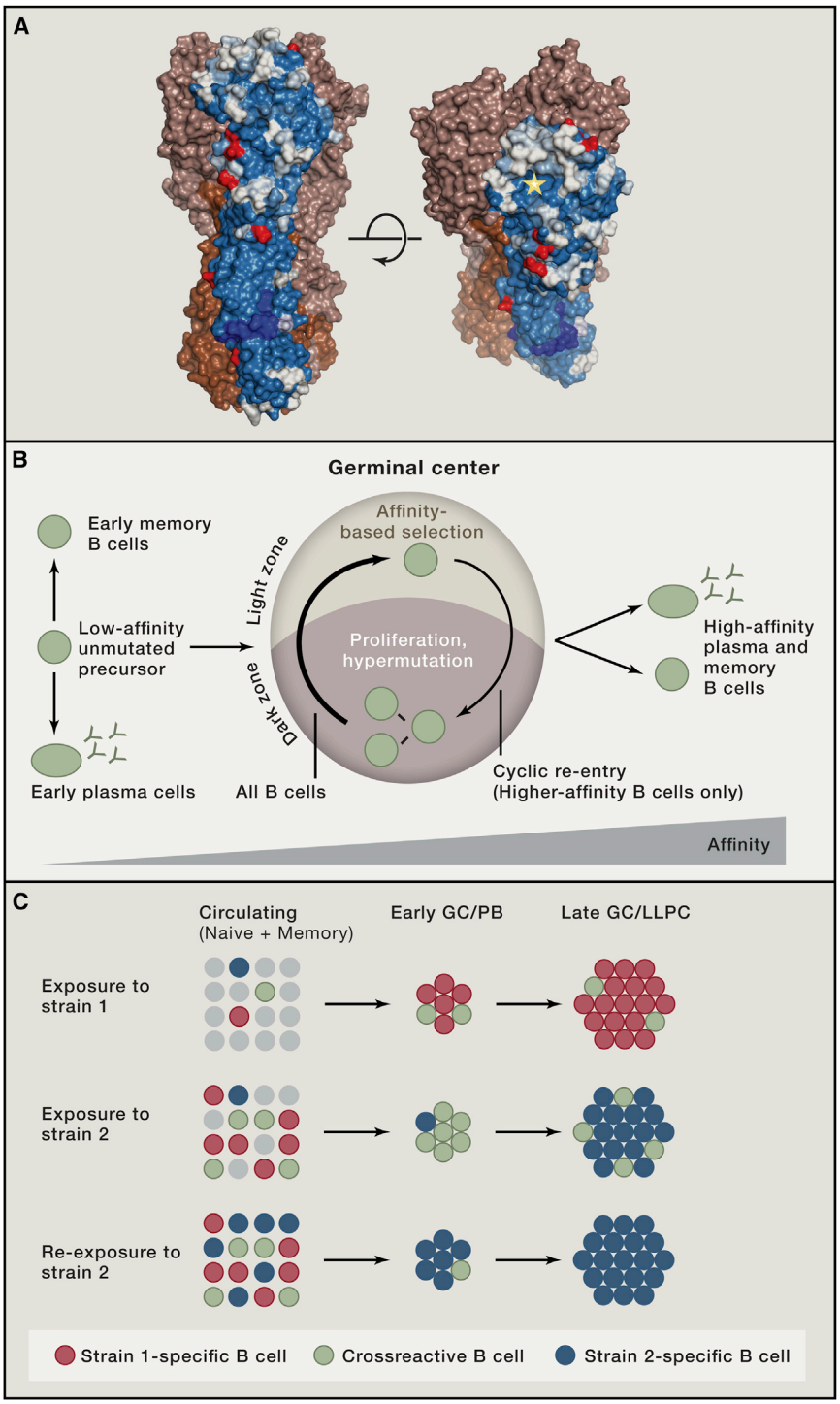


Figure 1. The Germinal Center Response to Influenza

(A) Residue conservation among seasonal H1 HA isolates (1975–2005). Conservation is shown for one monomer on a scale from blue (most conserved) to white (most variable). Red residues indicate common glycosylation sites. The RBS is marked with a star. Image courtesy of Stephen C. Harrison.

(B) Overview of affinity maturation in the GC. Cyclic migration of B cells between light and dark zones drives affinity maturation. Prior to GC entry and upon positive selection, B cells can differentiate into the PC or memory fates.

(C) Proposed model for re-establishment of immunodominance to strain-specific epitopes. Top: exposure to Strain 1 of influenza generates a response mostly focused on immunodominant, Strain-1-specific epitopes (red) but also induces a subdominant cross-reactive response (green). Middle: exposure to a divergent Strain 2 will initially reactivate cross-reactive memory cells (green) from the response to Strain 1, generating a "broad" response, but will also prime Strain-2-specific clones de novo, which will eventually outcompete the cross-reactive clones. Bottom: re-exposure to Strain 2 will preferentially recall Strain-2-specific clones, reinstating immunodominance.

The cues that trigger B cells to choose between cyclic re-entry and differentiation into PCs or memory B cells are unknown. High affinity for antigen appears to be a pre-requisite for PC differentiation and/or survival (Goodnow et al., 2010). However, because PC differentiation occurs after clonal expansion, diversion of part of a clone into the PC fate does not preclude further diversification of the same expanded clone in GCs. Also under scrutiny is the relative propensity of memory B cells to re-enter GCs for further diversification, rather than exclusively differentiating into secondary PC (McHeyzer-Williams et al., 2015). The ability to sequentially diversify the same clone over multiple responses is likely to be crucial to eliciting a broad response to influenza.

Selection of High-Affinity Mutants in the GC

Two models for how affinity-based selection operates in GCs are traditionally proposed. The first, and simplest, centers on antigen-driven signaling through the B cell receptor (BCR, comprising surface Ig, Igα, and Igβ) as the direct driver of selection. In this model, Ig with highest affinity for antigen will bind more strongly to immune complexes deposited on LZ FDCs, which triggers their return to the DZ and further proliferation (Victora and Nussenzweig, 2012). A recent development is the debate over whether BCR signaling is even active in GC B cells undergoing selection in the LZ and the role that inhibition of BCR signaling by Fc receptors might play in this process (Espéli et al., 2012; Khalil et al., 2012).

The second model of selection proposes that, rather than competing for direct signals from antigen, GC B cells compete

for limiting amounts of Tfh cell help. Here, the primary role of the BCR in selection is to trigger endocytosis: B cells acquire and present antigen in proportion to the affinity of their Ig. This maps Ig affinity—a B cell intrinsic property—onto surface peptide-MHC (pMHC) density—a feature that can be distinguished by Tfh cells. Several aspects of this model have been validated experimentally, and forcing interaction of GC B cells with Tfh is the only experimental approach so far that has been successful in triggering positive selection of GC B cells in vivo (Victora and Nussenzweig, 2012).

These two models are closely interrelated and thus not mutually exclusive. For example, strong inhibition by Fc receptors could serve to blunt BCR signaling so that B cells rely more heavily on T cells for selection. On the other hand, signals from T cells could potentially relieve Fc-mediated repression, allowing for productive BCR signals only in selected cells.

Establishing the mechanism of GC-positive selection can have important consequences to our understanding of how broadly protective antibodies develop. A recent report by Wang et al. (2015) has shown that levels of antibodies with sialylated Fcs in trivalent influenza vaccine (TIV) recipients 7 days after vaccination predicted the affinity of the anti-HA response 2 weeks later. Vaccination of mice with sialylated (versus non-sialylated) HA immune complexes generated antibodies capable of heterosubtypic protection in an in vivo challenge model. The authors traced this effect back to the upregulation of inhibitory Fc receptor FcγRIIB in GC B cells by sialylated Fcs, which would increase the threshold for BCR-driven selection, altering the affinity and/or specificity of the ensuing response. On the other hand, T cell priming with plasmid DNA encoding H1 prior to immunization with seasonal vaccine also increased the frequency of cross-reactive antibodies directed to the HA stem (Wei et al., 2010). While the mechanistic basis for this subversion of immunodominance is not clear, a likely scenario is that increased CD4 T cell priming may have led to relaxed interclonal competition between B cells before and within GCs because T cell help became less limiting.

Clonal Diversity and Immunodominance in the Antibody Response

The naive B cell repertoire comprises a large number of distinct V(D)J rearrangements, each expressed by only one or a few cells that proliferate to form clones upon antigenic exposure. B cells with low or even undetectable affinity for HA are capable of being recruited into GCs (Lingwood et al., 2012).

Competition between B cell clones (*inter*clonal competition) somewhat limits the access of lower-affinity B cells to the early GC. This clonality is further restricted in mature GCs by combined competition between clones and among SHM variants of the same clone (*intra*clonal competition) (Eisen, 2014; Victora and Nussenzweig, 2012). Competition is thought to lead to progressive loss of clonal diversity in the responding population. Thus, only a fraction of B cells remains in the immune response long enough to acquire the somatic mutations required to confer high affinity, leading to immunodominance. The immune system must therefore tune competition to the right level to balance affinity and diversity—if too stringent, average population affinity will increase fast but at the expense of diversity, and if too lax, diversity will remain high, but affinity will increase only slowly.

Immunodominance appears to be a key factor in preventing the emergence of broadly neutralizing influenza antibodies. Antibodies against epitopes that are conserved between different HA variants, such as the HA stem or the RBS, are underrepresented when compared to antibodies to more variable regions on the HA globular head. Potential reasons for this are that antibodies that bind these conserved epitopes require particular amino acid sequence elements (Lee and Wilson, 2015; Schmidt et al., 2015) and that conserved regions represent a relatively small or inaccessible portion of the HA surface (Figure 1A). Conversely, epitopes in the more variable regions of HA that are permissive to antigenic drift are more abundant, more accessible on the intact virion, and can be targeted in a multitude of ways. Evolutionary pressure on the virus may have led to the development of these variable but immunodominant epitopes as decoys, thus protecting conserved sites.

When exposed to a novel influenza strain for the first time, conserved epitopes are the only ones to which memory B cells exist. Thus, novel influenza strains can activate memory B cells that are cross-reactive to conserved epitopes, even predominantly generating a broad response (Wrammert et al., 2011; Ellebedy et al., 2014). However, re-exposure to a novel strain will shift the response predominantly toward antibodies on the globular head, reinstating its immunodominance (Ellebedy et al., 2014). We propose that such immunodominance is due in large part to GC (and potentially pre-GC) selection steering the antibody response away from conserved but subdominant epitopes toward more immunodominant ones (Figure 1C). Two factors that could contribute to re-establishing immunodominance are a greater potential of epitopes on the variable portion of HA to drive affinity maturation and incomplete conservation of cross-reactive epitopes between variant influenza strains. Thus, cross-reactive memory B cells may have sufficient affinity to become PCs and re-enter GCs upon exposure to a novel strain but could nonetheless be outcompeted in the course of the response by primary B cell clones undergoing de novo affinity maturation toward drifted but more immunodominant epitopes (Figure 1C).

The importance of GC selection for immunodominance is illustrated by a recent experiment in mice. Repeated administration of a low dose of the immune-suppressant rapamycin during influenza immunization abolished the GC response, which surprisingly was followed by increased resistance to heterosubtypic challenge and a change in the HA epitopes targeted by the resulting antibodies (Keating et al., 2013). The mechanistic reasons for this shift are unclear but are likely related to relaxed competition in the absence of a GC response. This observation suggests that immunodominance of certain regions of HA over others is not set in stone and can potentially be overcome by optimizing vaccination strategies to skew interclonal competition.

Approaches for Vaccination

Universal vaccination to influenza would require an antibody response that not only neutralizes all existent strains but also from which no variant can escape by mutation. Epidemiological evidence suggests that responses of such type can be elicited. For example, the broadly protective responses of humans to

the 2009 pandemic H1N1 strain may have caused the eradication of the previous H1N1 lineage that had infected humans for 91 years since the 1918 pandemic but no longer circulates (Krammer and Palese, 2015). Epidemiological studies aimed at identifying individuals who are completely immune to one or more influenza subtypes may help determine the required features of a universally protective immune response in a manner similar to what has been achieved in recent years by studying HIV-infected individuals (Klein et al., 2013).

A recent study by Schmidt et al. (2015) provides a glimpse of what a universally protective response might look like. In one individual, a series of clonally unrelated antibodies were found that bind to the conserved RBS pocket from different angles. In this case, mutations in the rim of the pocket, which normally render RBS antibodies ineffective against antigenic drift, are effective in preventing neutralization by only one or a few of these antibodies, but never all of them. Thus, perhaps a ''team'' of neutralizing antibodies may be able perform a function that would be impossible for any single bNAb.

Several studies in the literature have suggested strategies to elicit broadly neutralizing responses (Krammer and Palese, 2015). These follow along two broad lines: the first aims at developing immunization with a variety of natural or engineered antigens designed to force the immune system to focus on cross-reactive epitopes. These approaches include simultaneous or sequential immunization with different natural HA proteins, truncated (e.g., stem only) or chimeric (e.g., conserved stem, ''exotic'' head) HA variants, and/or viruses of varied subtype. The rationale is to attempt to overcome immunodominance by either eliminating strain-specific immunodominant epitopes or providing a competitive advantage to B cell clones that recognize epitopes common to multiple divergent HAs. Two recent reports highlight the promise of such strategies for inducing cross-reactive antibodies in multiple animal models (Impagliazzo et al., 2015; Yassine et al., 2015).

The second set of approaches relies on using standard antigens while manipulating the rules of selection in the antibody response. Some of these were discussed above (immunization with immune complexes, rapamycin treatment, and DNA priming). Strategies based on increasing Tfh help have been particularly of interest, given the great emphasis on Tfh cells as the ''judges'' of GC selection (Victora and Nussenzweig, 2012). A question that remains unsolved is what effect changing Tfh numbers has on selection: while fewer Tfh cells may promote stronger competition and therefore maximize the rate of affinity maturation, more Tfh may be desirable to maximize the size, quantity, and duration of GCs and perhaps allow for the appearance and maintenance of subdominant B cell clones. Evidence that adjuvants such as MF59 can expand the breadth of epitopes targeted by humans to include more of the conserved epitopes provides proof-of-principle evidence that manipulating the immune response can lead to increased clonal diversity (Del Giudice and Rappuoli, 2015).

Once a broadly protective response can be achieved by vaccination, a final issue that will need to be addressed is the longevity of the broadly protective response. That is, can an established broadly protective response resist challenge by immunodominant responses to drifting or non-protective epitopes eventually? Further understanding of the basic biology of recall responses and maintenance of long-lived PC will be required to address these issues.

ACKNOWLEDGMENTS

We thank S. Harrison, M. Nussenzweig, and F. Krammer for critical reading of our manuscript. G.D.V. is supported by NIH grant 5DP5OD012146. P.C.W. is supported by NIH grants 1P01AI097092-03, 2U19AI057266-11, U19AI109946, 1P01AI097092, and NIAID Center of Excellence for Influenza Research and Surveillance #HHSN272201400005C.

REFERENCES

Del Giudice, G., and Rappuoli, R. (2015). Curr. Top. Microbiol. Immunol. *386*, 151–180.

Eisen, H.N. (2014). Cancer Immunol. Res. *2*, 381–392.

Ellebedy, A.H., Krammer, F., Li, G.M., Miller, M.S., Chiu, C., Wrammert, J., Chang, C.Y., Davis, C.W., McCausland, M., Elbein, R., et al. (2014). Proc. Natl. Acad. Sci. USA *111*, 13133–13138.

Espéli, M., Clatworthy, M.R., Bökers, S., Lawlor, K.E., Cutler, A.J., Köntgen, F., Lyons, P.A., and Smith, K.G. (2012). J. Exp. Med. *209*, 2307–2319.

Goodnow, C.C., Vinuesa, C.G., Randall, K.L., Mackay, F., and Brink, R. (2010). Nat. Immunol. *11*, 681–688.

Impagliazzo, A., Milder, F., Kuipers, H., Wagner, M., Zhu, X., Hoffman, R.M., van Meersbergen, R., Huizingh, J., Wanningen, P., Verspuij, J., et al. (2015). Science *349*, 1301–1306.

Keating, R., Hertz, T., Wehenkel, M., Harris, T.L., Edwards, B.A., McClaren, J.L., Brown, S.A., Surman, S., Wilson, Z.S., Bradley, P., et al. (2013). Nat. Immunol. *14*, 1266–1276.

Khalil, A.M., Cambier, J.C., and Shlomchik, M.J. (2012). Science *336*, 1178–1181.

Klein, F., Mouquet, H., Dosenovic, P., Scheid, J.F., Scharf, L., and Nussenzweig, M.C. (2013). Science *341*, 1199–1204.

Krammer, F., and Palese, P. (2015). Nat. Rev. Drug Discov. *14*, 167–182.

Lee, P.S., and Wilson, I.A. (2015). Curr. Top. Microbiol. Immunol. *386*, 323–341.

Lingwood, D., McTamney, P.M., Yassine, H.M., Whittle, J.R., Guo, X., Boyington, J.C., Wei, C.J., and Nabel, G.J. (2012). Nature *489*, 566–570.

McHeyzer-Williams, L.J., Milpied, P.J., Okitsu, S.L., and McHeyzer-Williams, M.G. (2015). Nat. Immunol. *16*, 296–305.

Pappas, L., Foglierini, M., Piccoli, L., Kallewaard, N.L., Turrini, F., Silacci, C., Fernandez-Rodriguez, B., Agatic, G., Giacchetto-Sasselli, I., Pellicciotta, G., et al. (2014). Nature *516*, 418–422.

Schmidt, A.G., Therkelsen, M.D., Stewart, S., Kepler, T.B., Liao, H.X., Moody, M.A., Haynes, B.F., and Harrison, S.C. (2015). Cell *161*, 1026–1034.

Victora, G.D., and Nussenzweig, M.C. (2012). Annu. Rev. Immunol. *30*, 429–457.

Wang, T.T., Maamary, J., Tan, G.S., Bournazos, S., Davis, C.W., Krammer, F., Schlesinger, S.J., Palese, P., Ahmed, R., and Ravetch, J.V. (2015). Cell *162*, 160–169.

Wei, C.J., Boyington, J.C., McTamney, P.M., Kong, W.P., Pearce, M.B., Xu, L., Andersen, H., Rao, S., Tumpey, T.M., Yang, Z.Y., and Nabel, G.J. (2010). Science *329*, 1060–1064.

Wrammert, J., Koutsonanos, D., Li, G.M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W.I., et al. (2011). J. Exp. Med. *208*, 181–193.

Yassine, H.M., Boyington, J.C., McTamney, P.M., Wei, C.J., Kanekiyo, M., Kong, W.P., Gallagher, J.R., Wang, L., Zhang, Y., Joyce, M.G., et al. (2015). Nat. Med. *21*, 1065–1070.

Leading Edge
Perspective

Cell

A Relay Race on the Evolutionary Adaptation Spectrum

Avihu H. Yona,[1,2] Idan Frumkin,[3] and Yitzhak Pilpel[3,*]
[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
*Correspondence: pilpel@weizmann.ac.il
http://dx.doi.org/10.1016/j.cell.2015.10.005

Adaptation is the process in which organisms improve their fitness by changing their phenotype using genetic or non-genetic mechanisms. The adaptation toolbox consists of varied molecular and genetic means that we posit span an almost continuous "adaptation spectrum." Different adaptations are characterized by the time needed for organisms to attain them and by their duration. We suggest that organisms often adapt by progressing the adaptation spectrum, starting with rapidly attained physiological and epigenetic adaptations and culminating with slower long-lasting genetic ones. A tantalizing possibility is that earlier adaptations facilitate realization of later ones.

When challenged by new conditions, organisms adapt by changing their phenotype to improve fitness. The adaptation toolbox consists of varied molecular and genetic means: physiological acclimation, epigenetic changes, structural re-arrangements of the genome, and changes in the DNA sequence. Physiological responses, such as gene expression changes, are often the first to emerge upon environmental changes. Yet, although physiological adaptations may confer selective advantage, they are not actively amplified, memorized, or propagated over many generations. On the next level are epigenetic adaptations, which are distinct from the physiological adaptations, as they can have varying degrees of self-perpetuation over time, and they may occur at the DNA and chromatin, RNA, and even protein. As such, they constitute a molecular "memory." A next level on the spectrum is that of DNA copy-number adaptations, which include segmental DNA duplications/deletions that may range from specific genes to whole chromosomes. These are relatively labile genetic changes, although they do not involve changes in the actual nucleotide sequence of the genome. Lastly, genomic mutations represent the ultimate level of adaptation in which specific changes are stored and inherited relatively faithfully for prolonged periods. The diverse adaptations along the spectrum differ by several important attributes: the time needed for the adaptation to be attained at the individual organism level, the time until the adaptation becomes frequent in the population, the duration through which the adaptation can be sustained beyond the presence of the external selective pressure and when it is relieved, and the faithfulness and accuracy at which the adaptation is propagated across generations (Figure 1). When a challenge persists for longer durations, early adaptations that have been obtained may be subsequently replaced by a more durable adaptation. Indeed, it is often observed that adaptations at the various levels may facilitate one another, e.g., transcription changes can induce chromatin-based modifications (Henikoff and Shilatifard, 2011) and chromosome aneuploidy facilitates mutations in the DNA (Sheltzer et al., 2011). Although adapting organisms need not necessarily move linearly and uni-directionally along this "adaptation spectrum," the effective timescales of the different adaptations may dictate a tendency to move along the spectrum, from the short-lived physiological changes toward the long-lasting genetic ones. For example, a recent study on malaria discovered that *P. falciparum*-acquired drug resistance is a step-wise adaptation process. A non-genetic adaptation to the drug precedes duplications and mutations of the gene that confers the drug resistance (Herman et al., 2014). Along this line, we would like to hypothesize that organisms can perform a "relay race" on the adaptation spectrum.

Below, we discuss adaptations at each level of the spectrum—the context in which they occur and their typical timescales. Further we highlight cases of adaptation that may support the "relay race" notion, as some adaptions happen sequentially, each paving the way for the next to occur.

Physiological Adaptations

When stressed, organisms often acclimate quickly by a series of physiological responses. For example, in the yeast *S. cerevisiae*, a significant portion of the transcriptome changes in response to diverse environmental stressors such as extreme temperature, pH, salinity, and various drugs. These transcriptional plasticity responses are often temporary, as fast relaxation of the response is observed within minutes or hours, even as the stress prevails (Causton et al., 2001; Gasch et al., 2000; Shalem et al., 2008). What is the nature of this response? Genes that are needed to cope with the stress are often induced, e.g., heat-shock chaperones and anti-oxidants enzymes, while genes needed to sustain growth under optimal conditions like ribosomal genes are repressed. The durability of gene expression changes can vary and in some cases can persist across cellular generations. For example, when yeast cells are switched from glucose into galactose, they upregulate the galactose genes (Zacharioudakis et al., 2007). Yet, if switched back into glucose for one generation and then again into galactose, the expression response will be faster than at first encounter with galactose, suggesting a memory of the previous exposure to galactose. Even when grown for up to seven generations away from galactose, cells still "remember" the previous galactose experience.

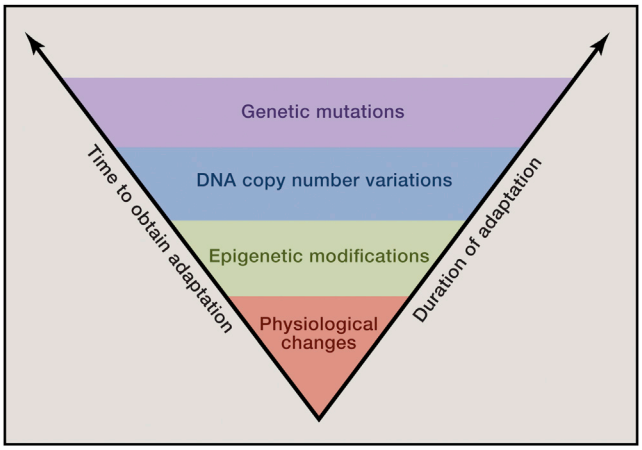


Figure 1. The Different Levels of the Adaptation Spectrum
Two timescales characterize the different adaptation levels: the time needed to acquire the adaptation (left axis) and the time along which the adaptation can be maintained in the absence of the condition that originally required the adaptation. Physiological adaptations consist of changes in current biochemical homeostasis, and therefore their duration depends on the lifetime of those biomolecules that underlie the adaptation, like mRNAs, proteins, etc. Epigenetic modifications typically occur within the same generation that experience the trigger, yet their duration depends on the type of epigenetic mechanism, e.g., DNA methylation, chromatin modification, prions, etc. (reviewed by Rando and Verstrepen, 2007). Genomic duplications include segmental duplications and aneuploidy as a result of chromosomal mis-segregation during cell cycle that result in crude changes in gene expression of the duplicated region. Genetic mutations are functional changes in the coding sequence or regulatory regions of genes that alter their function or expression.

Where and how is this memory stored? It was previously believed that the memory is implemented by nuclear factors that determine rate of transcription re-activation upon re-encounter with galactose (Brickner et al., 2007; Kundu et al., 2007). Yet, a later study clearly showed that the memory is stored in the cytoplasm, in the form of a signaling protein (Zacharioudakis et al., 2007; Ptashne, 2008). It is thus suggested that dilution of the protein in every cell division limits the durability of this memory to seven generations. More recent work on similar "phenotypic memory" showed a memory of up to ten generations when *E. coli* cells were subjected to rapidly alternating carbon sources. This memory mechanism, termed "response memory," appeared to be a hysteretic behavior in which gene expression persists after removal of its external inducer, and this enhances adaptation when environments fluctuate over short timescales (Lambert and Kussell, 2014).

In an environment that changes in a predictive manner, gene expression programs were found to encode "anticipation" of the subsequent environmental changes so that genes are expressed prior to the occurrence of the stimulus that normally activates them (Brunke and Hube, 2014; Mitchell et al., 2009; Tagkopoulos et al., 2008). But what if the challenges are unfamiliar? Response to an unforeseen challenge may require a dedicated strategy. In one study in yeast, a gene that is essential under the applied conditions was placed under a promoter that precludes its expression (Stern et al., 2007). Cells were thus trapped in a situation in which they must express a gene, which they possess, though in an inaccessible regulatory form. After approximately ten generations, a solution appears to have been found, as the population restored the ability to grow. The nature of this solution remains largely unknown. A potentially useful hint appears to be that, when genome-wide transcription was monitored, it was found to be different in each repetition of the experiment, suggesting that the cell's strategy might be to deliberately introduce noise to their expression program such that each cell will gamble on a potentially unique solution. In that respect, it could be appreciated that, like genetic mutations, which are predominantly neutral, gene expression changes might be neutral too (Koonin, 2007). Yet, whether noisy expression is the solution to the unforeseen challenge in this case is still an open question.

Rapid physiological reaction to a challenge is common among cells in the population and appears to be a first line of adaptation, which is mostly based on a hard-wired reaction to stimuli. This reaction is considered adaptive, as it not only improves the fitness under the current occurrence of the challenge, but as mentioned above, it might also improve the ability of the organism to cope with immediately subsequent occurrences of this challenge. In our context, physiological adaptation is defined by a lack of ability to actively perpetuate a memory. Nonetheless, changes in gene expression may serve as a substrate for downstream epigenetic modifications that can prolong their effect.

Epigenetic Adaptations

In this section, we distinguish physiological adaptations from the next modes of adaptation that feature active mechanisms for memory propagation across generations. Despite the controversy over the diverse definitions of epigenetics (Riddihough and Zahn, 2010), it is probably within the consensus that they involve some active mechanisms to perpetuate a memory across (cellular or even organismal) generations. As we will discuss, this epigenetic memory can be implemented at any level of the Central Dogma.

Inheritance by DNA Methylation and Chromatin Modifications

Apart from the nucleotide sequence itself, information on the DNA can be dynamically modified at two prime levels that constitute a major form of "epigenetics." One prime source of epigenetic information is implemented by covalent modification of DNA bases; the other is implemented by histones. DNA methylation of CpG dinucleotide in promoters with CpG islands is generally considered transcriptionally repressive (Cedar and Bergman, 2009). Due to the palindromic nature of CpG di-nucleotides, methylation of the C residue in the parental DNA strain can be easily restored in the new strand after cell division by recognition and methylation of hemi-methylated sites. Yet, the capacity to inherit epigenetic changes across organismal generations, e.g., in animals, is limited since it is *mostly* erased in the early embryo and then re-established in each individual (Smith et al., 2014).

DNA methylation is not the only epigenetic change that occurs on chromatin. Histone modifications, which occur in an impressive diversity of chemical forms and types, are long known to be associated with different states of transcription (Jenuwein and Allis, 2001). The "histone code" hypothesis asserts that some of the many modifications that take place on histones affect, either positively or negatively, transcription levels. However,

the issue is highly controversial (Henikoff and Shilatifard, 2011), as the main evidence for association between a particular chromatin modification and transcription activity level is often correlative rather than causal. Thus, the alternative to the histone code hypothesis is that certain marks on chromatin *result from*, rather than *cause*, a transcriptionally active or repressed state (Henikoff and Shilatifard, 2011). The accumulation of evidence in favor of each of the two directions may suggest a reconciled reality in which transcription state determines certain histone modifications, of which some can, in turn, affect transcription. Assume conservatively that transcription activation of a gene was regulated by a conventional transcription factor and that this change has consequently affected some histone marks in the vicinity of the regulated gene. These and other histone changes may sustain and perpetuate further the initial transcriptional activation. In other words, the mutual effect of transcription activity and histone marks could serve as memory loop with improved self-perpetuation capacity that transmits a purely physiological transcription change into the longer enduring epigenetic level. One demonstration of the effect of chromatin regulation was observed in an experiment that confronted flies with an unfamiliar challenge (a toxin), for which they were armed with a defense mechanism yet without a suitable regulatory program. Upon the first encounter with the toxin, flies had to suppress chromatin remodelers, the Polycomb genes, in order to activate the defense gene. This change appears to have led to the de-repression of developmental regulators in the affected organ (Stern et al., 2012), and some of the developmental alterations were epigenetically inherited by subsequent generations of unchallenged offspring. The possibility that histone marks are transferred across generations remains an open issue (Moazed, 2011). Nonetheless, recent indications from fission yeast show that chromatin marks can be inherited across many cell generations, independently of DNA sequence, DNA methylation, or RNA interference. Thus, histone marks constitute epigenetic information that can be perpetuated long after the removal of the initiating trigger (Audergon et al., 2015; Ragunathan et al., 2014).

RNA Inheritance

RNA can also transmit epigenetic information between generations. In *C. elegans*, dsRNA-mediated silencing has been shown to produce heritable responses (Fire et al., 1998). Recently, it has been further demonstrated that this nematode utilizes heritable RNAi responses to cope with environmental stresses. In one case, RNAi response was shown to be adaptive by silencing an infectious viral genome (Rechavi et al., 2011). In addition to viruses, heritable small RNAs serve to ward off other genomic parasites such as transposons (Ashe et al., 2012; Shirayama et al., 2012). In another case, RNA inheritance enabled memory of an environmental challenge even when no foreign DNA is incorporated: it was shown that RNAi response can be inherited following a developmental arrest caused by starvation (Rechavi et al., 2014). In this context, the effect also increased the longevity of the progeny by targeting genes with a role in nutrition. Importantly, it was found that this induced gene silencing is transmitted in a non-Mendelian manner that is not dependent on a DNA template but, rather, on an RNA-dependent RNA polymerase, which replicates the RNA to a sufficiently high level that overcomes the dilution effect across generations. In addition, small RNAs specifically induce the production of new small RNAs that spread also to nearby sequences (Sapetschnig et al., 2015). Notably, inheritance of small RNAs is dependent on specific factors, which are required for RNAi inheritance, but not for RNAi per se (Buckley et al., 2012).

In summary, RNA has been shown to propagate memory via different types of self-reinforcing epigenetic loops. These diverse classes of non-coding RNAs emerge as key regulators of gene expression typically by modifying chromatin structure and silence transcription (Holoch and Moazed, 2015).

Protein-Based Inheritance

Prions constitute a unique mechanism to perpetuate protein-based phenotypic changes. Unlike most proteins, prions can assume more than one stable conformation, in which the prionic conformation can serve as an auto-catalyst that can convert other conformations to the prion conformation (DeArmond and Prusiner, 2003). Importantly, prions can be acquired from the environment, e.g., through the diet, such as in the case of the Prion Protein Mad Cow Disease, or in response to other environmental changes (Lindquist, 1996). Such is the case of the translation terminator *SUP35* in yeast. In its non-prionic form, this protein serves as a release factor, needed for the proper translation termination of the ribosome at STOP codons. Yet, in its prion version, this protein aggregates and becomes less effective and accurate in terminating translation (Shorter and Lindquist, 2005). The outcome is therefore an extension of the polypeptide beyond the canonical STOP codon in a mechanism that might allow proteins to be extended with some stochasticity and potentially result in a population with enhanced phenotypic diversity (Halfmann et al., 2012). Like the above-mentioned stochastic transcriptome response and DNA methylation, this mechanism too can be activated upon stress (Halfmann and Lindquist, 2010), thus rapidly disseminating non-genetic diversity when diversity might be most needed. Yet, the feature that makes prion-based response truly exciting is its self-perpetuating nature. The autocatalytic tendency of the aggregation form appears to act as an epigenetic memory that perpetuates through generation and generates non-genetic diversity upon which natural selection can act. More recent work on another yeast prion, [GAR+], demonstrated how a prion can become adaptive by allowing cells to utilize more diverse metabolic capacities. Induced by bacteria, the [GAR+] prion state allows yeast to switch from purely fermenting glucose into a more versatile state that allows the simultaneous exploitation of diverse carbon sources (Jarosz et al., 2014a). Importantly, fitness of [GAR+] cells was found to be higher in low-glucose environments compared to cells in which the protein is in its non-prion state (Jarosz et al., 2014b).

In summary, epigenetic adaptation consists of a rich set of mechanisms that provide fascinating opportunities for organisms to rapidly disseminate variability in populations, long before genetic changes begin to fixate. But nonetheless, they are not heritable to the degree that genetic changes are. In the relay race context, there is a mutual effect between transcription activity and histone/DNA marks. As for the effect of epigenetics on later levels of the spectrum, it seems that the epigenetic architecture of genes also affects their chances of acquiring duplications or mutations (discussed below).

Adaptation by Changes in DNA Copy Number

As mentioned above, physiological and epigenetic adaptations are often carried out by changes in gene expression. Changes in DNA copy number are, in fact, another way to alter gene expression, yet this mode of adaptation is fundamentally distinct from the mechanisms mentioned above like transcription-factor-mediated changes. Genomic copy-number changes scale from single genes to aneuploidy (defined here as copy number change of whole chromosomes or parts of them). For most genes, a change in copy number results in altered mRNA levels as well as altered protein levels. This correlation between copy number and expression has been demonstrated in various organisms, including yeast (Dephoure et al., 2014; Pavelka et al., 2010a; Springer et al., 2010; Torres et al., 2007), plants (Huettel et al., 2008), mice (Kahlem et al., 2004; Lyle et al., 2004), and humans (Gao et al., 2007; Henrichsen et al., 2009; Stingele et al., 2012; Tsafrir et al., 2006; Williams et al., 2008). Therefore, DNA copy-number changes can be adaptive under selective pressures: when elevated expression is beneficial, extra copies can be acquired; conversely, when lower expression is beneficial, genomic copies can be lost. For example, the copy number of the human salivary amylase gene (*AMY1*) is positively correlated with the production level of salivary amylase protein, and populations with high-starch diets have more *AMY1* copies than those with traditionally low-starch diets (Perry et al., 2007).

When a higher expression of a specific gene is under selection, any genomic duplication that contains this gene has the potential to be adaptive. The most precise duplication would be of a small locus containing the gene in need. Yet, genomic adaptations are assumed to occur randomly, and thus the larger the duplicated region is, the higher the chances are for it to include the needed gene. For example, parallel *E. coli* populations evolved under limiting lactulose (a lactose isomer) showed duplication-based adaptations that varied in length. Although all duplication included the lactose permease (*lacY*), the shortest duplication covered 18 nearby genes and the largest consisted of up to 74 genes (Zhong et al., 2004). Notably, larger duplications come with a cost, as they contain many irrelevant genes whose expression is altered too. This altered expression of a large number of genes simultaneously imposes a significant burden on the cell (Bonney et al., 2015; Tang and Amon, 2013). Focusing on large copy-number variations like segmental aneuploidy or whole-chromosome aneuploidy (referred together as aneuploidy), it is important to note that despite their substantial cost, they have unique advantages and characteristics that distinguish them from the other forms of adaptation in the spectrum, as discussed below.

Aneuploidy as a Highly Accessible Evolutionary Solution

Aneuploidy is caused by mis-segregation of homologous chromosomes during cell division, and estimates indicate occurrence of 1:10,000 cell cycles in yeast and up to 1% in mammalian tissues (Knouse et al., 2014; Thompson and Compton, 2008; Zhu et al., 2014). Given these frequencies, populations of cells may constantly contain a variety of aneuploid cells that may be utilized as a resource for adaptation when facing a new challenge. Indeed, analysis of the yeast gene deletion library revealed that, in ~8% of the strains, deletion of a gene led to aneuploidy (Hughes et al., 2000). Interestingly, in some of the observed aneuploidies, the duplicated chromosome was found to harbor a close homolog of the deleted gene. In another study, a causal connection between aneuploidy and drug resistance was shown. The fungal pathogen *C. albicans* repeatedly acquired chromosome 5 aneuploidy in response to an antifungal drug exposure (Selmecki et al., 2008). The major mechanism by which duplication of chromosome 5 confers increased drug resistance is by amplifying two genes located in the duplicated chromosome: *ERG11* (encoding the drug target) and *TAC1* (encoding a transcriptional regulator of drug efflux pumps). Another yeast study showed that *S. cerevisiae* that have been evolved for ~200 generations under sulfate-limited conditions exhibited genomic duplications of regions that harbored the *SUL1* gene, which encodes a high-affinity sulfate transporter (Gresham et al., 2008). The rapid fixation of duplication-based adaptations mentioned above can be mainly attributed to their high occurrence in genomes and to the fact that duplications amplify many genes concurrently. This makes genomic duplications a highly accessible local maximum in the fitness landscape, whereas other adaptations are more complex and thus require longer evolutionary time to be acquired. An interesting hypothesis is that evolution acts to organize related genes on the same chromosome, perhaps even in proximity within the chromosome, so that duplications would elevate these genes together, with relatively fewer unrelated "hitchhiker" genes (Janga et al., 2008).

The Reversible Nature of Copy-Number-Based Adaptations

Aneuploidy-based adaptations are rapidly gained in evolution upon stress, but how reversible are such adaptations when the selection pressure is removed? The antifungal drug resistance that was facilitated by aneuploidy was shown to be reversible, as the extra chromosome was eliminated upon removal of the drug (Selmecki et al., 2006). In another study, yeast cells that were artificially selected for high expression of a single gene showed two types of distinct adaptations: duplication of large genomic regions (that contain the gene under selection) and *trans*-acting mutations. When selection was removed, only populations that adapted by aneuploidy could rapidly revert to base level (Rosin et al., 2012). This illustrates that adaptations based on duplications can serve as an "easy come easy go" adaptation, as when the stress is relieved, the costly duplication is driven out of the population much faster compared to sequence-based adaptations.

The Effectiveness of Aneuploidy for Acute and Abrupt Stresses

Genomic duplications appear to provide a rescue when a selective pressure is introduced in an abrupt manner, but would they appear also when stress is slowly aggravating? A recent lab evolution study (Yona et al., 2012) directly tested the effect of the stress regime, abrupt versus gradual, on the type of the selected evolutionary solution. This work demonstrated that, when yeast cells were abruptly shifted from 30°C to 39°C, where they evolved for some 500 generations, adaptation was repeatedly achieved by duplication of chromosome 3. Yet, populations that were evolved under a different heat regime in which temperature increased gradually (from 30°C to 39°C by +1°C increments every 50 generations) did not adapt by genomic

duplications but, rather, by sequence mutations. This suggests that, due to the high cost of aneuploidy, it is not an efficient response unless the selective pressure is acute and abrupt. Curiously, genome sequencing of the evolved populations shows that populations evolved under the abrupt heat-shock regime duplicated chromosome 3 but did not fixate any point mutation, while the populations that evolved under the gradual heat regime fixated 8–12 point mutations. It is tempting to speculate that this result could prove to be more general—that is, other conditions that select for aneuploidy under abrupt stress would select for changes other than aneuploidy when the same stress is applied gradually.

Genomic Duplications as a Transient Solution that Can Be Refined by Focal Adaptations

When the chromosome 3 aneuploid yeast were further evolved for additional >1,000 generations under high temperature, the extra copy of chromosome 3 was lost and replaced by a series of point mutations (Yona et al., 2012). The state of chromosome duplication thus appears as a transient step, an evolutionary "stepping stone." A similar evolutionary dynamic was observed in an *E. coli* study that showed how cells with impaired lac operon adapted first by multiple duplications of the impaired genes, as means to increase expression (Hendrickson et al., 2002). This amplification not only enabled lactose utilization, but also made the lac-operon hypermutable, as any additional copies increase the chances of finding beneficial mutations in this locus. Indeed, shortly after the gene amplification, one of the duplicated copies acquired mutations that restored a high functional level that led to the subsequent elimination of the other low-functional copies (Hendrickson et al., 2002). Interestingly, this dynamics may also be relevant to pathogens that adapt drug resistance in the clinics. A recent study on clinical isolates of *C. albicans* suggested that, in some isolates, aneuploidies may have an important role as an intermediate adaptation that subsequently gives rise to more stable adaptive genotypes that confer drug resistance (Ford et al., 2015). To conclude, it seems that prolonged evolution can solve the paradox of aneuploidy (Pavelka et al., 2010b; Sheltzer and Amon, 2011): Under normal conditions, selection purges fitness-lowering aneuploidy. Yet, under abrupt stresses, beneficial aneuploidy is selected because it confers higher survivability and proliferation to enable expansion of the effective population that can further search the fitness landscape for more optimal and slowly acquired solutions. Thus, aneuploidy appears as a transient step along the adaptation spectrum that facilitates a path to the next stage: sequence-based mutations.

Adaptation of the Genetic Sequence: Mutations on the Spectrum

The adaptation that is most commonly identified with evolution is genetic mutations. Mutations make the long-lasting adaptations, and unlike the previous stages along the adaptation spectrum, they can change not only expression regulatory regimes, but also actual protein sequence and function.

Adaptation through duplications or mutations could serve as alternative evolutionary strategies, but which genes go through which track? A recent study found that the answer could be in promoters' architecture. A careful analysis of nucleosome arrangement within promoters revealed a surprising deviation in some genes with the classical nucleosome-free region (NFR) architecture. It was found that genes whose expression is typically required at a certain level, like housekeeping genes, tend to have an NFR and low transcriptional plasticity i.e., low transcriptional variation across conditions (Tirosh and Barkai, 2008). It was further shown experimentally that the expression level of these NFR genes features low evolvability, i.e., their expression level is relatively insensitive to promoter mutations (Hornung et al., 2012). In contrast, genes that require a more dynamic transcription, like stress genes, typically lack an NFR, and their transcription is highly plastic and can be effectively altered by mutations (Hornung et al., 2012; Tirosh and Barkai, 2008). In a follow-up lab-evolution experiment (Rosin et al., 2012), yeast cells were put under short-term selection for higher expression of specific genes that were deliberately chosen to represent either cases of classical NFR or its absence. Interestingly, when the gene under selection had an NFR, selection toward higher expression was achieved by a large segmental or even whole-chromosomal duplications of regions that harbor the gene, presumably because of the low ability of these genes to elevate transcription by mutations. Conversely, for genes with no NFR, higher expression was achieved by mutations and with no duplications. This notion that nucleosome architecture of genes can create a bias that affects downstream adaptations on the spectrum (duplication versus mutations) highlights another relay race dynamic that connects between the epigenetic and genetic levels.

Here, we focus on genetic mutations only in the context of the adaptation spectrum and discuss some of the unique features of this adaptation mode. According to the "modern synthesis" of genetics and evolution, mutations are seeded at random in genomes, irrespective of environmental conditions or potential phenotypic effects. Nonetheless, there is some evidence suggesting that genetic mutations can be more accessible in challenging conditions and perhaps also in specific genomic regions.

Stress-Induced Mutagenesis and Transcription-Coupled Mutations

Diverse environmental stresses induce a higher rate of mutations (or lower efficiency of repair) (Gentile et al., 2011; Giraud et al., 2001; Loh et al., 2010; Oliver et al., 2000). In that respect, just like noise in gene expression or epigenetic noise in DNA methylation, mutations introduce diversity in populations especially under stress, when higher mutation rate may be beneficial. Subsequently, cells carrying mutations that confer higher fitness will prevail and improve the population's fitness. It is possible that stress-induced mutagenesis might not be adaptive, and it occurs simply due to the fact that, under stress, many processes in the cell are less accurate. Yet, indications on the precise control of error-prone DNA synthesis and some theoretical considerations on mutagenesis point toward the adaptive nature of increased mutagenesis under stress (Ishii et al., 1989; Lynch, 2010; Sniegowski et al., 2000). Mutagenesis, or DNA repair following mutagenesis, changes not only over time, but also spatially along the genome (Schuster-Böckler and Lehner, 2012; Supek and Lehner, 2015). Non-uniform distribution of mutations in different genomic regions suggests another potential

feature that may improve the efficient utilization of genomic mutations as a means of adaptation. What is truly interesting in our context is the process known as "transcription-coupled mutagenesis," by which the rate of mutation is elevated in proportion to transcription rate (Jinks-Robertson and Bhagwat, 2014). Thus, if transcription regulation allows "reading" different parts of the genome in different environments, transcription-coupled mutagenesis and related processes (Howan et al., 2012) may allow "(re)-writing," i.e., mutagenizing different parts of the genome at different rates under specific conditions. While previous chapters of this Perspective described protein and RNA inheritance of a Lamarckian nature, stress-induced mutagenesis and transcription-coupled mutagenesis both introduce some Lamarckian changes to the DNA level as well. Indeed, natural evolution seems now more Lamarckian than we thought until recently (Koonin and Wolf, 2009).

Phenotypic Mutations and Their Interaction with Genetic Mutations

Another means to diversify the proteome and the transcriptome before DNA mutations start to appear is known by the collective term "phenotypic mutations," representing errors in transcription and translation (Bürger et al., 2006). Of all processes in the Central Dogma, DNA replication often occurs with the highest fidelity. While error rate of DNA replication is typically between 10^{-8} and 10^{-10} per base per cell cycle, the error rate of transcription and translation, per nucleotide or per amino acid, could be up to a million times higher (Gordon et al., 2009; Meyerovich et al., 2010; Pan, 2013). Conceivably, phenotypic mutations should not propagate between generations, as the lifetimes of RNAs and proteins are typically short, even compared to the generation time of unicellular species. However, phenotypic mutations can propagate by triggering transcription network loop, by interaction with genetic mutations, or even by assimilation into the genome. A study on the lac operon of *E. coli*, which comprises an autocatalytic positive-feedback loop, has demonstrated a heritable epigenetic switch (Gordon et al., 2009, 2013). In this study, transcription infidelity generated a mutated lac repressor with reduced ability to repress the lac operon. This led the lac operon to be more sensitive to lactose, i.e., the operon could be induced by lower concentration of lactose. In such a case, a positive-feedback loop of the induced operon state of transcription is propagated across generations. Another interesting dynamic that involves phenotypic mutations lies within the interaction between phenotypic and genetic mutations. The "look-ahead mutations" concept (Whitehead et al., 2008) is a putative form of adaptation in which phenotypic mutations facilitate the fixation of high-complexity *genetic* mutations. Imagine, for example, that a new disulfide bridge between two cysteine residues is beneficial in a specific protein. Creating a new disulfide bond requires two genetic mutation events, in each of which a non-cysteine is converted into a cysteine. The evolutionary catch is that the organism's fitness does not increase before the *later* of the two mutation events occurs, i.e., since the first mutation alone would confer no (or even negative) fitness gain, it has lower chances of existing in the population. The "look-ahead mutation" is a theoretical scenario in which one of the two mutations is a phenotypic mutation, while the other is a genetic one. Under certain realistic quantitative assumptions (regarding error rates, etc.), it was shown that, indeed, a hybrid of phenotypic and genetic double mutant could emerge and be sustained. For example, a cell in the population that carries the first genetic mutation can obtain a second phenotypic mutation with partial functionality that will increase its fitness and, thus, its fraction in the population. Following that, the phenotypic mutation can be replaced with a fully functional genetic counterpart that will further increase the fitness (see Figure 3 in review by Koonin, 2012). In that respect, the phenotypic mutation may serve as an intermediate evolutionary "stepping stone" that can be rapidly attained and later replaced.

Phenotypic mutations might also be assimilated directly into the genome—for example, via the process of reverse transcription (RT). RT is not only used by retro viruses; it also occurs in cellular life, and it might act on genes in addition to retrotransposons (Cordaux and Batzer, 2009). This mechanism appears to be relevant in cancer, as a recent study found intron-less versions of human genes in cancerous genomes. These newly formed retro genes most likely result from reverse transcription of certain transcripts that are acquired somatically during cancer development (Cooke et al., 2014). Given the high rate of transcription errors, RT could serve as a potential evolvability mechanism by which phenotypic mutations become genetic.

Genetic Redundancy: The Longest-Lasting Adaptation?

We have proposed an adaptation spectrum that culminates in hard-wired genetic changes. Such changes are indeed stably "memorized" by genomes. However, even for a genetic adaptation, memory is not guaranteed to be indefinite. Mutational drift is certainly possible especially in periods when environmental conditions no longer necessitate the previously adaptive change. In this respect, what is more stable than a stable genetic change? Perhaps two stable genetic changes. Indeed, biological redundancy is prevalent in many genomes, and it is often suggested to provide a "fail-safe" mechanism, or backup: if one of two redundant genes is mutated, the other can still perform the lost function, albeit often upon a change in expression program (DeLuna et al., 2010; Kafri et al., 2009). The evolutionary stability of redundant genetic states is not trivial and can be sustained only under certain conditions (Nowak et al., 1997). Nonetheless, it is expected that, if the selective pressure that necessitated the adaptation is removed, the genetic adaptation will still be sustained, provided that the process of "neofunctionalization" (He and Zhang, 2005) has not yet taken place.

The processes that can occur along the adaptation spectrum have largely been described within the conceptual framework of one cell's lineage, and indeed we have been deploying the word adaptation in the context of an individual cell's (or organism's) improved fitness. However, to translate to evolutionarily meaningful changes and, indeed, to meet the more commonly recognized meaning of the term "adaptation," the described beneficial changes within a cell/organism need to ultimately result in changes at the population level. Our thinking on this is as follows. Consider an environmental stress that necessitates high expression of a particular gene. Cells in the population that highly express this gene, either in response to the stress or even prior to its occurrence, will have a temporary advantage over other cells. This higher expression may be achieved by stochastic differences at the physiological or epigenetic level. These

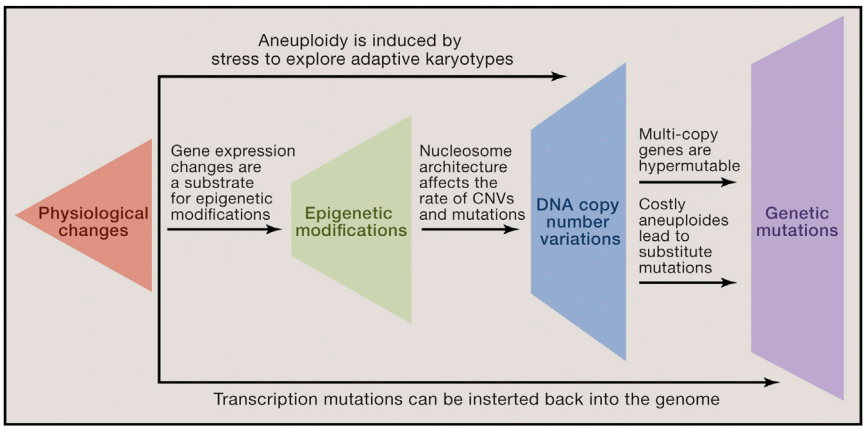


Figure 2. The Interplay between Different Modes of Adaptation on the Spectrum
Illustration of directional interactions between adaptation modes, where early adaptations with short persistence affect later, more durable, adaptations. Short descriptions of such interactions (arrows) demonstrate how later adaptations are more focused according to the trajectory set by earlier adaptations. We suggest that all modes of adaptation progress as in a relay race to optimize the whole process of organismal adaptation.

cells that exhibit higher expression have a dual advantage in the population: first, they benefit from a higher fitness (as long as the stress persists); second, due to relay-race dynamics, their lineages may have higher chances of propagating the original short-term acclimation into a more sustainable genetic adaptation. Such cell lineages will thus have an advantage both in terms of cell number and in terms of an increased per-cell probability of further adaptations. The outcome could be that, by the time the stress is relieved and the population returns back to its "ground state," the physiological acclimation has already been propagated to the genetic level and it now prevails in the population.

A Relay-Race Cascade within the Adaptation Spectrum

We have delineated an evolutionary adaptation spectrum along which organisms may progress as they adapt to a new challenge. We have taken the risk of generalization in suggesting a stereotypical order of exploration along the spectrum, starting from the physiological adaptation and gradually moving toward the genetic, though deviations from this simple-minded search strategy could certainly be envisaged. In this last section, we discuss the possibility that, in some cases, realization of a given stage along the spectrum could facilitate the progression into a next stage, as in a relay race (Figure 2).

Starting with the physiological level, changes in gene expression contribute to the first line of adaptation; nonetheless, they may also actively set in motion later modes of adaptation. It has been widely suggested that gene expression affects the epigenetic "chromatin landscape" (Henikoff and Shilatifard, 2011). For example, expression of a non-coding RNA induces epigenetic silencing of ribosomal genes by interaction with their promoter (Schmitz et al., 2010). Physiological changes in gene expression may also induce genomic duplications like aneuploidy. Interestingly, the higher rates of aneuploidy observed after stress are connected to the activity of the chaperone Hsp90 (Chen et al., 2012). In our context, Hsp90 may represent a relay-race baton that mediates between these two modes of adaptation. On top of being a stress-response gene, Hsp90 has an evolutionarily conserved role in the kinetochore assembly (Niikura et al., 2006), and therefore it also has a role in aneuploidy formation. Therefore, physiological modulation of Hsp90 under stress is suggested to facilitate a search for adaptations by diversifying the karyotype. Further down the spectrum, the classical role of Hsp90 can also be discussed in the context of the relay race. As a chaperone, Hsp90 was shown to act as an "evolutionary capacitor"—when active, it appears to mask the effect of mutations on the phenotype, and when repressed, those cryptic variations can be exposed (Rohner et al., 2013; Rutherford and Lindquist, 1998). In that respect, this protein serves as a baton in the relay race, mapping its own expression onto the effect of sequence mutation on the phenotype. Finally, physiological adaptations might also have an effect on genetic mutations. RNA transcripts (which typically carry more mutations) can serve as a template for DNA repair by homologous recombination of the original genomic sequence from which they were transcribed (Keskin et al., 2014) or, as mentioned above, by actual integration of a cDNA reverse-transcription product into the genome, as shown in cancer (Cooke et al., 2014). Such processes may facilitate rapid evolution of currently expressed genes, with a useful bias in favor of highly expressed genes, as they produce more RNA copies that can be reinserted into the genome.

Further down the spectrum, another relay-race dynamic occurs when genomic duplications promote subsequent genetic mutations. Large duplications (like aneuploidy) not only increase the mutation rate (Sheltzer et al., 2011), but also favor mutations that are related to the duplicated region. First, increased copy number increases the probability of a mutation in one of the duplicated copies. Second, the excessive cost of large duplications may affect mutations by favoring those who can replace the duplication. Any emerging mutation(s) that can replace the duplication-based adaptation is reinforced by an additional fitness increase at the magnitude of the cost that was saved. This added advantage increases the likelihood of such mutations to fixate faster than other mutations that are not related to the initial duplication. In this way, there is a bias on subsequent mutations to cope with the selective pressure that led to the initial duplication-based adaptation.

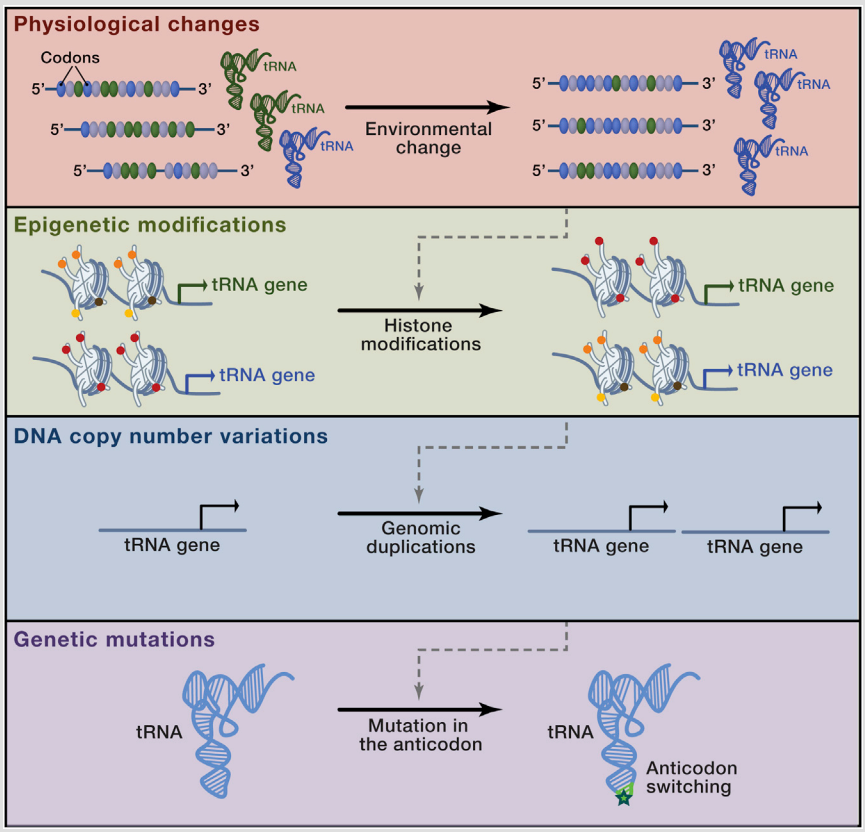
We present the example of translational optimization with which progression along the adaptation spectrum can be conceptualized with a specific set of cellular processes in Box 1.

A major topic that has not been discussed here in great length is that of cancer. Cancer is an evolutionary process, and as such, it may exploit different adaptions along the spectrum, as well as the relay race between them. To begin with, the cancerous transcriptome is known to be radically reprogrammed (c.f., Segal

Box 1. The Adaptation Spectrum of the Translation Process: A Test Case

Many of the cellular resources, i.e., energy and raw materials, are devoted to ensure adequate translation of proteins. Consequently, translation optimization is a major driving force in evolution. One of the factors that governs translation optimization is the balance between supply and demand i.e., between the tRNA pool to the codons used by currently expressed mRNAs. Perturbations in the supply-to-demand balance emerge when a new environment requires the expression of different proteins with a different codon usage (Gingold et al., 2012) or because the availability of some tRNAs is altered (Dittmar et al., 2005; Pavon-Eternod et al., 2013; Wiltrout et al., 2012). Examining how cells restore translational balance reveals many of the stages along the adaptation spectrum.

Measuring expression of the tRNA pool under diverse conditions shows physiological response in which distinct tRNA types are up/downregulated (Gingold et al., 2014). On the next level, epigenetics is becoming increasingly appreciated in this context, as histone modifications around tRNA genes change dynamically in response to cells' conditions (Barski et al., 2010; Gingold et al., 2014; Oler et al., 2010), e.g., in cancer. Further on the spectrum, duplication of tRNA genes appears to provide elevated expression from certain tRNAs that are in high demand, and indeed, many of the tRNA genes occur in multiple-copy families that change their relative sizes in evolution (Man and Pilpel, 2007). Therefore, it will be interesting to ask whether some of the chromosome gains and losses observed in cancer correspondingly increase or decrease tRNAs' availability to support the progression on cancer. Next, mutations within tRNAs were also found to be adaptive, presumably in response to change in the demand-to-supply ratio. When a tRNA gene is artificially deleted from the yeast genome, another tRNA gene with a different anticodon but of the same amino acid evolutionarily "responds" with a mutation that converts its anticodon to that of the deleted one (Yona et al., 2013). Such "anticodon switching" was subsequently found to be very prevalent in the natural evolution of species (Rogers and Griffiths-Jones, 2014; Yona et al., 2013). In each of these cases, we do not know whether the anticodon-switching mutation was preceded by earlier transcriptional/epigenetic changes, yet it is tempting to speculate that such physiological changes may have constituted an intermediate solution to the challenge before it was solved genetically. Further, the tRNA pool probably also realizes the last stage of the adaptation spectrum, i.e., that of genetic redundancy by compensation over mutated tRNAs (Bloom-Ackermann et al., 2014). Thus, partial redundancy among tRNAs may act to increase evolutionary stability on one hand and to facilitate evolutionary plasticity of the tRNA pool on the other hand.



The green and blue ovals represent codons that correspond to the anticodons of the green and blue tRNAs. Prior to an environmental change (upper-left), there is a high usage (translation demand) of a certain codon (green oval) compared to another codon (blue oval), and the tRNA levels (translational supply) match accordingly. An environmental shift (upper-right) may result in a physiological change both at the codon usage (now, the blue codon is in higher demand, because mRNAs that are enriched in the codon are induced), and tRNA levels adjust correspondingly. The higher expression of the blue tRNA could then be propagated into the epigenetic level, e.g., through changes in activation or repression-associated histone mark in the tRNA genes' vicinity. Such changes in the tRNA pool may be further implemented by changes to tRNA gene copy number. Finally, more copies of the same tRNA gene may increase its probability of acquiring both functional and regulatory mutations, like anticodon switching. Dashed arrows (gray) connecting the different levels represent hypothesized relay race between the levels.

et al., 2004), and recent analyses of cancer epigenomes showed that DNA methylation is stochastic rather than precise (Landan et al., 2012; Landau et al., 2014). The question of whether these changes at the physiological and epigenetic levels are connected remains unknown. Further along the spectrum, aneuploidy is a hallmark of cancer, and despite the debate of whether it is a cause or a consequence of cancer (Sheltzer and Amon, 2011), aneuploidy is suggested to provide cancer with both higher mutation rate and with a faster means to change dosages of cancer-driving genes, i.e., upregulating expression of oncogenes or downregulating tumor-suppressor genes (reviewed by Gordon et al., 2012). Furthermore, since aneuploidy can

increase copy number of cancer-driving genes and mutation rates simultaneously, it may lead to a hypermutability effect of oncogenes that were duplicated. Taken together, it is intriguing to speculate that cancer cells might exploit the relay race notion proposed here in gaining more aggressive traits much faster.

In conclusion, we suggest that the distinct modes of adaptation have been optimized by evolution not only to perform their adaptive function, but also to interact with later modes of adaptation. In this way, the whole process of adaption can yield better results as the fitness landscape is being explored more efficiently according to the trajectory set by earlier adaptations.

ACKNOWLEDGMENTS

We thank the Minerva Foundation for the establishment of the Minerva Center for Live Emulation of Genome Evolution, as well as the European Research Council (ERC) for grant support. We thank the Human Frontier Science Program for supporting A.H.Y. and the Azrieli Foundation for the Azrieli Fellowship to I.F. Y.P. is the incumbent of the Ben May Professorial Chair at the Weizmann Institute of Science.

REFERENCES

Ashe, A., Sapetschnig, A., Weick, E.M., Mitchell, J., Bagijn, M.P., Cording, A.C., Doebley, A.L., Goldstein, L.D., Lehrbach, N.J., Le Pen, J., et al. (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. Cell *150*, 88–99.

Audergon, P.N.C.B., Catania, S., Kagansky, A., Tong, P., Shukla, M., Pidoux, A.L., and Allshire, R.C. (2015). Restricted epigenetic inheritance of H3K9 methylation. Science *348* (80-.), 132–135.

Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat. Struct. Mol. Biol. *17*, 629–634.

Bloom-Ackermann, Z., Navon, S., Gingold, H., Towers, R., Pilpel, Y., and Dahan, O. (2014). A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. PLoS Genet. *10*, e1004084.

Bonney, M.E., Moriya, H., and Amon, A. (2015). Aneuploid proliferation defects in yeast are not driven by copy number changes of a few dosage-sensitive genes. Genes Dev. *29*, 898–903.

Brickner, D.G., Cajigas, I., Fondufe-Mittendorf, Y., Ahmed, S., Lee, P.-C., Widom, J., and Brickner, J.H. (2007). H2A.Z-mediated localization of genes at the nuclear periphery confers epigenetic memory of previous transcriptional state. PLoS Biol. *5*, e81.

Brunke, S., and Hube, B. (2014). Adaptive prediction as a strategy in microbial infections. PLoS Pathog. *10*, e1004356.

Buckley, B.A., Burkhart, K.B., Gu, S.G., Spracklin, G., Kershner, A., Fritz, H., Kimble, J., Fire, A., and Kennedy, S. (2012). A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. Nature *489*, 447–451.

Bürger, R., Willensdorfer, M., and Nowak, M.A. (2006). Why are phenotypic mutation rates much higher than genotypic mutation rates? Genetics *172*, 197–206.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001). Remodeling of yeast genome expression in response to environmental changes. Mol. Biol. Cell *12*, 323–337.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. *10*, 295–304.

Chen, G., Bradford, W.D., Seidel, C.W., and Li, R. (2012). Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. Nature *482*, 246–250.

Cooke, S.L., Shlien, A., Marshall, J., Pipinikas, C.P., Martincorena, I., Tubio, J.M.C., Li, Y., Menzies, A., Mudie, L., Ramakrishna, M., et al.; ICGC Breast Cancer Group (2014). Processed pseudogenes acquired somatically during cancer development. Nat. Commun. *5*, 3644.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

DeArmond, S.J., and Prusiner, S.B. (2003). Perspectives on prion biology, prion disease pathogenesis, and pharmacologic approaches to treatment. Clin. Lab. Med. *23*, 1–41.

DeLuna, A., Springer, M., Kirschner, M.W., and Kishony, R. (2010). Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLoS Biol. *8*, e1000347.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., and Torres, E.M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. eLife *3*, e03023.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep. *6*, 151–157.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Ford, C.B., Funt, J.M., Abbey, D., Issi, L., Guiducci, C., Martinez, D.A., Delorey, T., Li, B.Y., White, T.C., Cuomo, C., et al. (2015). The evolution of drug resistance in clinical isolates of Candida albicans. eLife *4*, e00662.

Gao, C., Furge, K., Koeman, J., Dykema, K., Su, Y., Cutler, M.L., Werts, A., Haak, P., and Vande Woude, G.F. (2007). Chromosome instability, chromosome transcriptome, and clonal evolution of tumor cell populations. Proc. Natl. Acad. Sci. USA *104*, 8995–9000.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Gentile, C.F., Yu, S.-C., Serrano, S.A., Gerrish, P.J., and Sniegowski, P.D. (2011). Competition between high- and higher-mutating strains of Escherichia coli. Biol. Lett. *7*, 422–424.

Gingold, H., Dahan, O., and Pilpel, Y. (2012). Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res. *40*, 10053–10063.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sørensen, K.D., et al. (2014). A dual program for translation regulation in cellular proliferation and differentiation. Cell *158*, 1281–1292.

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., and Taddei, F. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science *291*, 2606–2608.

Gordon, A.J.E., Halliday, J.A., Blankschien, M.D., Burns, P.A., Yatagai, F., and Herman, C. (2009). Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. PLoS Biol. *7*, e44.

Gordon, D.J., Resio, B., and Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. Nat. Rev. Genet. *13*, 189–203.

Gordon, A.J.E., Satory, D., Halliday, J.A., and Herman, C. (2013). Heritable change caused by transient transcription errors. PLoS Genet. *9*, e1003595.

Gresham, D., Desai, M.M., Tucker, C.M., Jenq, H.T., Pai, D.A., Ward, A., DeSevo, C.G., Botstein, D., and Dunham, M.J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. *4*, e1000303.

Halfmann, R., and Lindquist, S. (2010). Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science *330*, 629–632.

Halfmann, R., Jarosz, D.F., Jones, S.K., Chang, A., Lancaster, A.K., and Lindquist, S. (2012). Prions are a common mechanism for phenotypic inheritance in wild yeasts. Nature *482*, 363–368.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics *169*, 1157–1164.

Hendrickson, H., Slechta, E.S., Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2002). Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. Proc. Natl. Acad. Sci. USA *99*, 2164–2169.

Henikoff, S., and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet. *27*, 389–396.

Henrichsen, C.N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. *41*, 424–429.

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. *16*, 71–84.

Hornung, G., Oren, M., and Barkai, N. (2012). Nucleosome organization affects the sensitivity of gene expression to promoter mutations. Mol. Cell *46*, 362–368.

Howan, K., Smith, A.J., Westblade, L.F., Joly, N., Grange, W., Zorman, S., Darst, S.A., Savery, N.J., and Strick, T.R. (2012). Initiation of transcription-coupled repair characterized at single-molecule resolution. Nature *490*, 431–434.

Huettel, B., Kreil, D.P., Matzke, M., and Matzke, A.J.M. (2008). Effects of aneuploidy on genome structure, expression, and interphase organization in Arabidopsis thaliana. PLoS Genet. *4*, e1000226.

Hughes, T.R., Roberts, C.J., Dai, H., Jones, A.R., Meyer, M.R., Slade, D., Burchard, J., Dow, S., Ward, T.R., Kidd, M.J., et al. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. Nat. Genet. *25*, 333–337.

Ishii, K., Matsuda, H., Iwasa, Y., and Sasaki, A. (1989). Evolutionarily stable mutation rate in a periodically changing environment. Genetics *121*, 163–174.

Janga, S.C., Collado-Vides, J., and Babu, M.M. (2008). Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. Proc. Natl. Acad. Sci. USA *105*, 15761–15766.

Jarosz, D.F., Brown, J.C., Walker, G.A., Datta, M.S., Ung, W.L., Lancaster, A.K., Rotem, A., Chang, A., Newby, G.A., Weitz, D.A., et al. (2014a). Cross-kingdom chemical communication drives a heritable, mutually beneficial prion-based transformation of metabolism. Cell *158*, 1083–1093.

Jarosz, D.F., Lancaster, A.K., Brown, J.C.S., and Lindquist, S. (2014b). An evolutionarily conserved prion-like element converts wild fungi from metabolic specialists to generalists. Cell *158*, 1072–1082.

Herman, J.D., Rice, D.P., Ribacke, U., Silterra, J., Deik, A.A., Moss, E.L., Broadbent, K.M., Neafsey, D.E., Desai, M.M., Clish, C.B., et al. (2014). A genomic and evolutionary approach reveals non-genetic drug resistance in malaria. Genome Biol. *15*, 511.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074–1080.

Jinks-Robertson, S., and Bhagwat, A.S. (2014). Transcription-associated mutagenesis. Annu. Rev. Genet. *48*, 341–359.

Kafri, R., Springer, M., and Pilpel, Y. (2009). Genetic redundancy: new tricks for old genes. Cell *136*, 389–392.

Kahlem, P., Sultan, M., Herwig, R., Steinfath, M., Balzereit, D., Eppens, B., Saran, N.G., Pletcher, M.T., South, S.T., Stetten, G., et al. (2004). Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. Genome Res. *14*, 1258–1267.

Keskin, H., Shen, Y., Huang, F., Patel, M., Yang, T., Ashley, K., Mazin, A.V., and Storici, F. (2014). Transcript-RNA-templated DNA recombination and repair. Nature *515*, 436–439.

Knouse, K.A., Wu, J., Whittaker, C.A., and Amon, A. (2014). Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proc. Natl. Acad. Sci. USA *111*, 13409–13414.

Koonin, E.V. (2007). Chance and necessity in cellular response to challenge. Mol. Syst. Biol. *3*, 107.

Koonin, E.V. (2012). Does the central dogma still stand? Biol. Direct *7*, 27.

Koonin, E.V., and Wolf, Y.I. (2009). Is evolution Darwinian or/and Lamarckian? Biol. Direct *4*, 42.

Kundu, S., Horn, P.J., and Peterson, C.L. (2007). SWI/SNF is required for transcriptional memory at the yeast GAL gene cluster. Genes Dev. *21*, 997–1004.

Lambert, G., and Kussell, E. (2014). Memory and fitness optimization of bacteria under fluctuating environments. PLoS Genet. *10*, e1004556.

Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A., et al. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat. Genet. *44*, 1207–1214.

Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., et al. (2014). Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell *26*, 813–825.

Lindquist, S. (1996). Mad cows meet mad yeast: the prion hypothesis. Mol. Psychiatry *1*, 376–379.

Loh, E., Salk, J.J., and Loeb, L.A. (2010). Optimization of DNA polymerase mutation rates during bacterial evolution. Proc. Natl. Acad. Sci. USA *107*, 1154–1159.

Lyle, R., Gehrig, C., Neergaard-Henrichsen, C., Deutsch, S., and Antonarakis, S.E. (2004). Gene expression from the aneuploid chromosome in a trisomy mouse model of down syndrome. Genome Res. *14*, 1268–1274.

Lynch, M. (2010). Evolution of the mutation rate. Trends Genet. *26*, 345–352.

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat. Genet. *39*, 415–421.

Meyerovich, M., Mamou, G., and Ben-Yehuda, S. (2010). Visualizing high error levels during gene expression in living bacterial cells. Proc. Natl. Acad. Sci. USA *107*, 11543–11548.

Mitchell, A., Romano, G.H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O., and Pilpel, Y. (2009). Adaptive prediction of environmental changes by microorganisms. Nature *460*, 220–224.

Moazed, D. (2011). Mechanisms for the inheritance of chromatin states. Cell *146*, 510–518.

Niikura, Y., Ohta, S., Vandenbeldt, K.J., Abdulle, R., McEwen, B.F., and Kitagawa, K. (2006). 17-AAG, an Hsp90 inhibitor, causes kinetochore defects: a novel mechanism by which 17-AAG inhibits cell proliferation. Oncogene *25*, 4133–4146.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature *388*, 167–171.

Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., Graves, B.J., and Cairns, B.R. (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nat. Struct. Mol. Biol. *17*, 620–628.

Oliver, A., Cantón, R., Campo, P., Baquero, F., and Blázquez, J. (2000). High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science *288*, 1251–1254.

Pan, T. (2013). Adaptive translation as a mechanism of stress response and adaptation. Annu. Rev. Genet. *47*, 121–137.

Pavelka, N., Rancati, G., Zhu, J., Bradford, W.D., Saraf, A., Florens, L., Sanderson, B.W., Hattem, G.L., and Li, R. (2010a). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature *468*, 321–325.

Pavelka, N., Rancati, G., and Li, R. (2010b). Dr Jekyll and Mr Hyde: role of aneuploidy in cellular adaptation and cancer. Curr. Opin. Cell Biol. *22*, 809–815.

Pavon-Eternod, M., Gomes, S., Rosner, M.R., and Pan, T. (2013). Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA

expression and increased proliferation in human epithelial cells. RNA *19*, 461–466.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Ptashne, M. (2008). Transcription: a mechanism for short-term memory. Curr. Biol. *18*, R25–R27.

Ragunathan, K., Jih, G., and Moazed, D. (2014). Epigenetic inheritance uncoupled from sequence-specific recruitment. Science (80-.), science.1258699 –.

Rando, O.J., and Verstrepen, K.J. (2007). Timescales of genetic and epigenetic inheritance. Cell *128*, 655–668.

Rechavi, O., Minevich, G., and Hobert, O. (2011). Transgenerational inheritance of an acquired small RNA-based antiviral response in C. elegans. Cell *147*, 1248–1256.

Rechavi, O., Houri-Ze'evi, L., Anava, S., Goh, W.S.S., Kerk, S.Y., Hannon, G.J., and Hobert, O. (2014). Starvation-induced transgenerational inheritance of small RNAs in C. elegans. Cell *158*, 277–287.

Riddihough, G., and Zahn, L.M. (2010). Epigenetics. What is epigenetics? Introduction. Science *330*, 611.

Rogers, H.H., and Griffiths-Jones, S. (2014). tRNA anticodon shifts in eukaryotic genomes. RNA *20*, 269–281.

Rohner, N., Jarosz, D.F., Kowalko, J.E., Yoshizawa, M., Jeffery, W.R., Borowsky, R.L., Lindquist, S., and Tabin, C.J. (2013). Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish. Science *342* (80-.), 1372–1375.

Rosin, D., Hornung, G., Tirosh, I., Gispan, A., and Barkai, N. (2012). Promoter nucleosome organization shapes the evolution of gene expression. PLoS Genet. *8*, e1002579.

Rutherford, S.L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. Nature *396*, 336–342.

Sapetschnig, A., Sarkies, P., Lehrbach, N.J., and Miska, E.A. (2015). Tertiary siRNAs mediate paramutation in C. elegans. PLoS Genet. *11*, e1005078.

Schmitz, K.-M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev. *24*, 2264–2269.

Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504–507.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. *36*, 1090–1098.

Selmecki, A., Forche, A., and Berman, J. (2006). Aneuploidy and isochromosome formation in drug-resistant Candida albicans. Science *313*, 367–370.

Selmecki, A., Gerami-Nejad, M., Paulson, C., Forche, A., and Berman, J. (2008). An isochromosome confers drug resistance in vivo by amplification of two genes, ERG11 and TAC1. Mol. Microbiol. *68*, 624–641.

Shalem, O., Dahan, O., Levo, M., Martinez, M.R., Furman, I., Segal, E., and Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. Mol. Syst. Biol. *4*, 223.

Sheltzer, J.M., and Amon, A. (2011). The aneuploidy paradox: costs and benefits of an incorrect karyotype. Trends Genet. *27*, 446–453.

Sheltzer, J.M., Blank, H.M., Pfau, S.J., Tange, Y., George, B.M., Humpton, T.J., Brito, I.L., Hiraoka, Y., Niwa, O., and Amon, A. (2011). Aneuploidy drives genomic instability in yeast. Science *333*, 1026–1030.

Shirayama, M., Seth, M., Lee, H.C., Gu, W., Ishidate, T., Conte, D., Jr., and Mello, C.C. (2012). piRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. Cell *150*, 65–77.

Shorter, J., and Lindquist, S. (2005). Prions as adaptive conduits of memory and inheritance. Nat. Rev. Genet. *6*, 435–450.

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. Nature *511*, 611–615.

Sniegowski, P.D., Gerrish, P.J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. BioEssays *22*, 1057–1066.

Springer, M., Weissman, J.S., and Kirschner, M.W. (2010). A general lack of compensation for gene dosage in yeast. Mol. Syst. Biol. *6*, 368.

Stern, S., Dror, T., Stolovicki, E., Brenner, N., and Braun, E. (2007). Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. Mol. Syst. Biol. *3*, 106.

Stern, S., Fridmann-Sirkis, Y., Braun, E., and Soen, Y. (2012). Epigenetically heritable alteration of fly development in response to toxic challenge. Cell Rep. *1*, 528–542.

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., and Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Mol. Syst. Biol. *8*, 608.

Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81–84.

Tagkopoulos, I., Liu, Y.-C., and Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. Science *320*, 1313–1317.

Tang, Y.-C., and Amon, A. (2013). Gene copy-number alterations: a cost-benefit analysis. Cell *152*, 394–405.

Thompson, S.L., and Compton, D.A. (2008). Examining the link between chromosomal instability and aneuploidy in human cells. J. Cell Biol. *180*, 665–672.

Tirosh, I., and Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. Genome Res. *18*, 1084–1091.

Torres, E.M., Sokolsky, T., Tucker, C.M., Chan, L.Y., Boselli, M., Dunham, M.J., and Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924.

Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R.F., Barany, F., et al. (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res. *66*, 2129–2137.

Whitehead, D.J., Wilke, C.O., Vernazobres, D., and Bornberg-Bauer, E. (2008). The look-ahead effect of phenotypic mutations. Biol. Direct *3*, 18.

Williams, B.R., Prabhu, V.R., Hunter, K.E., Glazier, C.M., Whittaker, C.A., Housman, D.E., and Amon, A. (2008). Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. Science *322*, 703–709.

Wiltrout, E., Goodenbour, J.M., Fréchin, M., and Pan, T. (2012). Misacylation of tRNA with methionine in Saccharomyces cerevisiae. Nucleic Acids Res. *40*, 10494–10506.

Yona, A.H., Manor, Y.S., Herbst, R.H., Romano, G.H., Mitchell, A., Kupiec, M., Pilpel, Y., and Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. Proc. Natl. Acad. Sci. USA *109*, 21010–21015.

Yona, A.H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J.D., Dahan, O., and Pilpel, Y. (2013). tRNA genes rapidly change in evolution to meet novel translational demands. eLife *2*, e01339.

Zacharioudakis, I., Gligoris, T., and Tzamarias, D. (2007). A yeast catabolic enzyme controls transcriptional memory. Curr. Biol. *17*, 2041–2046.

Zhong, S., Khodursky, A., Dykhuizen, D.E., and Dean, A.M. (2004). Evolutionary genomics of ecological specialization. Proc. Natl. Acad. Sci. USA *101*, 11719–11724.

Zhu, Y.O., Siegal, M.L., Hall, D.W., and Petrov, D.A. (2014). Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. USA *111*, E2310–E2318.

Cell

Leading Edge
Review

Mitochondrial ROS Signaling in Organismal Homeostasis

Gerald S. Shadel[1,2,3,*] and Tamas L. Horvath[3,4,5,*]
[1]Department of Pathology
[2]Department of Genetics
[3]Program in Integrative Cell Signaling and Neurobiology of Metabolism
[4]Section of Comparative Medicine
[5]Department of Neurobiology
Yale School of Medicine, New Haven CT 06520
*Correspondence: gerald.shadel@yale.edu (G.S.S.), tamas.horvath@yale.edu (T.L.H.)
http://dx.doi.org/10.1016/j.cell.2015.10.001

Generation, transformation, and utilization of organic molecules in support of cellular differentiation, growth, and maintenance are basic tenets that define life. In eukaryotes, mitochondrial oxygen consumption plays a central role in these processes. During the process of oxidative phosphorylation, mitochondria utilize oxygen to generate ATP from organic fuel molecules but in the process also produce reactive oxygen species (ROS). While ROS have long been appreciated for their damage-promoting, detrimental effects, there is now a greater understanding of their roles as signaling molecules. Here, we review mitochondrial ROS-mediated signaling pathways with an emphasis on how they are involved in various basal and adaptive physiological responses that control organismal homeostasis.

Mitochondria and Associated Homeostatic and Stress Signaling Pathways

Mitochondria are essential organelles present in all but a few mammalian cell types, where they perform multiple functions. They are the sites of the tricarboxylic acid (TCA) cycle and oxidative phosphorylation (OXPHOS), through which large amounts of ATP are generated using the electrochemical gradient generated across the inner of two membranes by the electron transport chain (ETC). However, their critical roles in metabolism go far beyond glucose oxidation via OXPHOS and include fatty acid and amino acid metabolism and biosynthesis of hormones, heme, and iron sulfur clusters. Furthermore, in addition to metabolism, mitochondria are involved in apoptosis, ion homeostasis, and innate immunity, with new roles in cell and organismal biology being discovered at an unprecedented rate.

Mitochondria are complex in composition, form, and function. Though often depicted as small round or oval structures, they are instead usually dynamic, branched networks that constantly fuse and divide under control of specific fission and fusion machineries (Mishra and Chan, 2014). Proteomic analyses indicate that mammalian mitochondria contain ~1,200 proteins, with the precise composition varying significantly between cell and tissue types (Calvo and Mootha, 2010). Thirteen of these proteins are encoded by the maternally inherited mitochondrial DNA (mtDNA) located in the matrix, while the rest are encoded by nuclear genes and targeted to the organelle by specific protein import pathways (Shadel and Clayton, 1997). Thus, mitochondrial biogenesis and homeostasis, including mtDNA expression and maintenance, are under strict control of nuclear gene expression programs (Scarpulla, 2008).

The overall status of mitochondria is constantly monitored, allowing their number, morphology, distribution, and activity to be modulated by developmental, physiological, and environmental cues. This requires bi-directional signaling pathways that meditate crosstalk between mitochondria and the nucleus. Pioneering studies in budding yeast revealed that mitochondrial dysfunction leads to so-called "retrograde signaling" events that result in adaptive changes in nuclear gene expression and metabolism mediated by specific transcription factors (Butow and Avadhani, 2004). Mitochondrial retrograde signaling pathways also exist in mammals and are now receiving considerable attention because they drive both beneficial and pathogenic adaptive responses.

Given their complicated nature, mitochondrial stress can manifest in many forms that elicit different stress signals. Reduced ETC/OXPHOS capacity can result in cellular energy deprivation (e.g., reduced ATP/energy charge), altered mitochondrial ROS (mtROS) production, or loss of mitochondrial membrane potential, with the precise outcome dictating the specific mitochondrial stress-signaling response (Butow and Avadhani, 2004; Sena and Chandel, 2012). Reduced mitochondrial protein import, improper assembly of large enzymatic complexes (e.g., OXPHOS and ribosomes), and altered chaperone activity can cause proteotoxic stress and mitochondrial unfolded protein responses (Haynes et al., 2013; Rugarli and Langer, 2012). As major sites of ROS production, mitochondria are also prone to oxidative damage and stress. Damage, mutation, or depletion of mtDNA causes distinct forms of mitochondrial stress and downstream signaling (Scheibye-Knudsen et al., 2015; West et al., 2015). Finally, altered morphology, dynamics, and distribution can lead to distinct forms of stress and are linked to

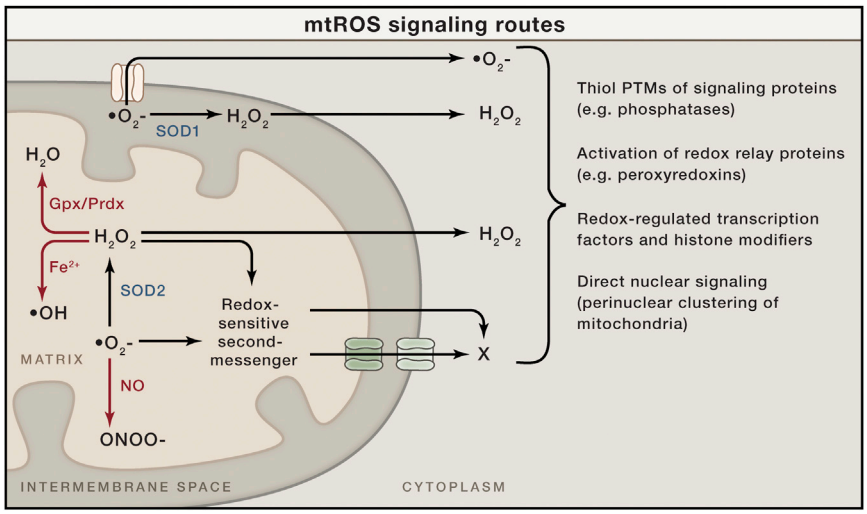


Figure 1. Mitochondrial ROS Signaling Basics

Superoxide ($\cdot O_2-$) is generated on both sides of the inner mitochondrial membrane and hence arises in the matrix or the intermembrane space (IMS). Superoxide can be converted to hydrogen peroxide (H_2O_2) by superoxide dismutase enzymes (SOD1 in the IMS or SOD2 in the matrix). The resulting hydrogen peroxide can cross membranes and enter the cytoplasm to promote redox signaling. Superoxide is not readily membrane permeable but may be released into the cytoplasm through specific outer membrane channels, as shown (see main text). In addition to signaling in the cytoplasm directly, both superoxide and hydrogen peroxide could, in principle, oxidize or modify other molecules in mitochondria that can be released/exposed to the cytoplasm to signal (redox-sensitive second messenger; X). These mitochondrial ROS (mtROS) can generate signaling responses and changes in nuclear gene expression in multiple ways (shown to the right). There are other fates of mtROS that would prevent signaling (or potentially enact other signaling and damage responses). For example, superoxide can react with nitric oxide (NO) to form peroxinitrite ($ONOO-$). This would prevent its conversion to hydrogen peroxide, could cause damage by the highly reactive peroxynitrite, and could potentially limit NO availability for its own type of signaling. Hydrogen peroxide can be eliminated enzymatically by glutathione peroxidase (Gpx) in the matrix or peroxiredoxins (Prdx) in the matrix and elsewhere in the cell. Peroxyredoxins can also promote redox signaling by promoting disulfide bond formation in target proteins. Finally, in the presence of transition metals, hydrogen peroxide can generate damaging hydroxyl radicals ($\cdot OH$).

mitochondrial turnover by autophagy or mitophagy (Labbé et al., 2014). Changes in these parameters and associated stress-signaling responses can occur downstream of physiological (e.g., nutrient limitations, substrate availability, exercise), environmental (exposure to drugs or toxins), and genetic cues. With regard to genetics, the importance of these pathways is underscored by the fact that inherited mitochondrial diseases are caused by mutations in genes encoding proteins involved in each of these processes that can be inherited maternally (mtDNA mutations) or in a Mendelian fashion (nuclear gene mutations) (Nunnari and Suomalainen, 2012). However, much remains to be learned about these stress pathways. For example, the specific sensors of different forms of mitochondrial stress and the cell and tissue specificity of the signaling responses remain largely unknown. Furthermore, the degree of crosstalk between different mitochondrial stress pathways and what determines whether they elicit beneficial or maladaptive responses are not clear.

Mitochondrial ROS Signaling

ROS are formed by one-electron transfers from a redox donor to molecular oxygen (O_2). This initially generates the anionic free-radical superoxide that can be converted to hydrogen peroxide by superoxide dismutase enzymes (Figure 1). Hydroxyl radical is another ROS that can be formed (e.g., by metal-catalyzed oxidation of hydrogen peroxide), but in this Review, "ROS" refers to superoxide and hydrogen peroxide unless otherwise noted. In mitochondria, the orderly flow of electrons down the mitochondrial ETC to complex IV results in their final deposition into molecular oxygen to form water. However, electrons can also react prematurely with oxygen at sites in the ETC to form superoxide/hydrogen peroxide (Murphy, 2009). Complexes I and III are often regarded as the major sites of mtROS production, but more recent studies indicate that at least ten other mitochondrial enzymes also contribute, including complex II (Quinlan et al., 2013). That different sites of mtROS production have distinct signaling roles and the primary production sites likely change under different physiological conditions is likely (Quinlan et al., 2013; Sena and Chandel, 2012).

That hydrogen peroxide has robust signaling roles in cells was elucidated through studies of receptor tyrosine kinase, growth-factor signaling that showed bursts of ROS production by NADPH oxidase (NOX) enzymes. A major mechanism at play in this scenario is the inactivation of redox-sensitive protein tyrosine phosphatases that normally downregulate these receptors (via dephosphorylation) by localized NOX-dependent production of hydrogen peroxide. Several paradigms emerge from these studies that are relevant to mitochondrial hydrogen peroxide acting as a signal (Finkel, 2012). First, many NOX enzymes produce extracellular superoxide that dismutates to hydrogen peroxide that is then is transported across the plasma membrane, perhaps in a regulated manner through aquaporin channels, to effect localized redox signaling. In a similar manner, superoxide produced in the mitochondrial inner-membrane space or matrix can be converted to hydrogen peroxide by SOD1 or SOD2, respectively, allowing it to diffuse into the cytoplasm to signal (Figure 1). Whether this involves free diffusion or facilitated diffusion through specific channels in mitochondrial membranes remains unclear. Second, the inactivation of phosphatases by hydrogen peroxide occurs through the modification of specific reactive thiol side chains (e.g., cysteine). It is now recognized that cysteine residues on many proteins can undergo a variety of redox-dependent modifications, including sequential oxidation (to sulfenic, sulfinic, and sulfonic acid), glutathiolation, and S-nitrosation (Go et al., 2015). Like phosphorylation, ubiquitination, and other post-translational modifications, these redox modifications can alter protein structure and function and be regulatory. Therefore, selective oxidation or modification of

redox-dependent thiols in regulatory proteins allows for intricate cellular redox-switch control mechanisms (Finkel, 2012; Go et al., 2015). Redox regulatory proteins that associate with or are otherwise selectively tuned to readout mitochondrial hydrogen peroxide production would provide a mechanism for mtROS signaling (Figure 1). Perinuclear clustering of mitochondria has also been postulated to be a mechanism for direct mitochondrial-nuclear signaling via mtROS (Al-Mehdi et al., 2012).

Superoxide is often summarily dismissed as a relevant signaling molecule because of its chemical properties. For example, unlike hydrogen peroxide, it is a negatively charged molecule and hence not able to easily diffuse across cell membranes, and it does not engage in protein cysteine oxidation reactions conducive to known redox-switch mechanisms of signaling (Winterbourn, 2008). However, as we will discuss, physiologically relevant, superoxide-mediated signaling does appear to exist that is distinct from hydrogen-peroxide-mediated signaling pathways. Although not "freely" diffusible, mitochondrial superoxide can be released from the intermembrane space into the cytoplasm through the voltage-dependent anion channel that spans the outer mitochondrial membrane (Han et al., 2003) (Figure 1). This includes superoxide that is generated in the intermembrane space by complex III, as well as that generated in the matrix by complex I (and potentially by other enzymes) (Lustgarten et al., 2012). However, the latter may require superoxide levels to cross a critical threshold (e.g., when antioxidant defenses are limiting). In the budding yeast, *S. cerevisiae*, mitochondrial superoxide is released into the cytoplasm by a specific isoform of the voltage-dependent anion channel (Por1p) or, in the absence of Por1p, through the TOM protein import complex (Budzińska et al., 2009). Thus, superoxide released from mitochondria can, in principle, participate directly in cytoplasmic signaling processes (Figure 1). Finally, it is possible that mitochondrial matrix superoxide signals to the cytoplasm via redox-sensitive second-messenger systems that have yet to be defined (Figure 1).

Mitochondrial ROS Signaling in Organismal Physiology: Lessons from Model Systems

There is now extensive evidence from the study of model organisms supporting an active role for mtROS signaling in organismal physiology and adaptive responses (Hamanaka and Chandel, 2010; Ristow and Zarse, 2010; Yun and Finkel, 2014). The tractability of these genetic model systems has led to elucidation of new molecular details and signaling pathways underlying these responses. In this regard, much has been learned in the context of aging and longevity studies, which will be highlighted here.

In the nematode worm, *C. elegans*, several pathways that extend lifespan involve increased mtROS production and signaling. These studies have called into question the "mitochondrial" and "free radical" theories of aging (at least as originally formulated) by implicating ROS as pro-longevity signals as opposed to damaging, pro-aging agents, as they've long been viewed. Ristow and colleagues broke new ground in this area by showing that reduced glucose availability leads to increased mitochondrial respiration and mtROS production that delays worm aging (Schulz et al., 2007). They and others have subsequently found that increased mtROS is a common downstream event in many conserved longevity-promoting interventions, which has led to the concept of "mitohormesis" (Ristow and Zarse, 2010; Yun and Finkel, 2014). Recent work in this area includes mtROS signaling in the anti-aging effects of reduced insulin/IGF signaling and D-glucosamine supplementation (Weimer et al., 2014; Zarse et al., 2012). Inhibition of the mitochondrial ETC by certain mutations or inactivation of mitochondrial SOD2 increases worm lifespan and has been causally linked to increased mtROS production (Dancy et al., 2014). Hekimi and colleagues have recently shown that this involves a unique form of activation of apoptotic signaling cascades to promote protective stress responses rather than apoptosis (Yee et al., 2014). Longevity-extending effects of mtROS in worms are also mediated by HIF1 and AMP kinase signaling and are linked to some degree to enhanced immunity (Hwang et al., 2014; Lee et al., 2010). Like in worms, mtROS signaling extends chronological lifespan in *S. cerevisiae*, which, in part, is how reduced TORC1 signaling mediates longevity in this organism (Bonawitz et al., 2007; Pan et al., 2011; Schroeder et al., 2013). Here, the mtROS signal activates the DNA-damage-sensing kinases, Tel1p and Rad35p (yeast orthologs of ATM and Chk2), leading to enhanced subtelomeric silencing via inactivation of Rph1p, a histone H3K36 demethylase of the jumonji family of enzymes (Schroeder et al., 2013). This response, vis-à-vis the mtROS-mediated, apoptotic-signaling response in worms discussed above (Yee et al., 2014), suggests that a new paradigm is emerging whereby canonical stress-response pathways (e.g., DNA repair and apoptosis) are utilized differentially to sense mtROS to elicit adaptive, homeostatic responses in addition to the emergency and cell death responses for which they were defined originally.

While the above discussion was limited largely to examples from worms and yeast, it is important to note that similar mtROS longevity pathways have been shown to operate in other invertebrates and in mice (Hekimi et al., 2011; Ristow and Schmeisser, 2011). Furthermore, these pathways are not limited to anti-aging responses. For example, mtROS signaling has also been implicated in other homeostatic pathways and processes, including wound healing (Xu and Chisholm, 2014), survival under hypoxia (Schieber and Chandel, 2014), intracellular pH homeostasis (Johnson et al., 2012), cell differentiation (Hamanaka and Chandel, 2010; Hamanaka et al., 2013; Tormos et al., 2011), and innate immunity (West et al., 2011). Accordingly, the remainder of this Review will be devoted to the role of mtROS in whole-body physiology, with the main focus on neuroendocrine control of systemic metabolism in mammals.

ROS Generation and Central Control of Whole-Body Metabolism

The amount of mtROS generated in metabolic processes depends on the fuel load and type (lipid, carbohydrate, protein), as well as the amount, composition, activity, and dynamics of mitochondria in the cell or tissue involved. Because ROS are de facto by-products of mitochondrial oxidative metabolism, it may not be surprising that studies have connected mtROS to neuroendocrine control of metabolism, including feeding

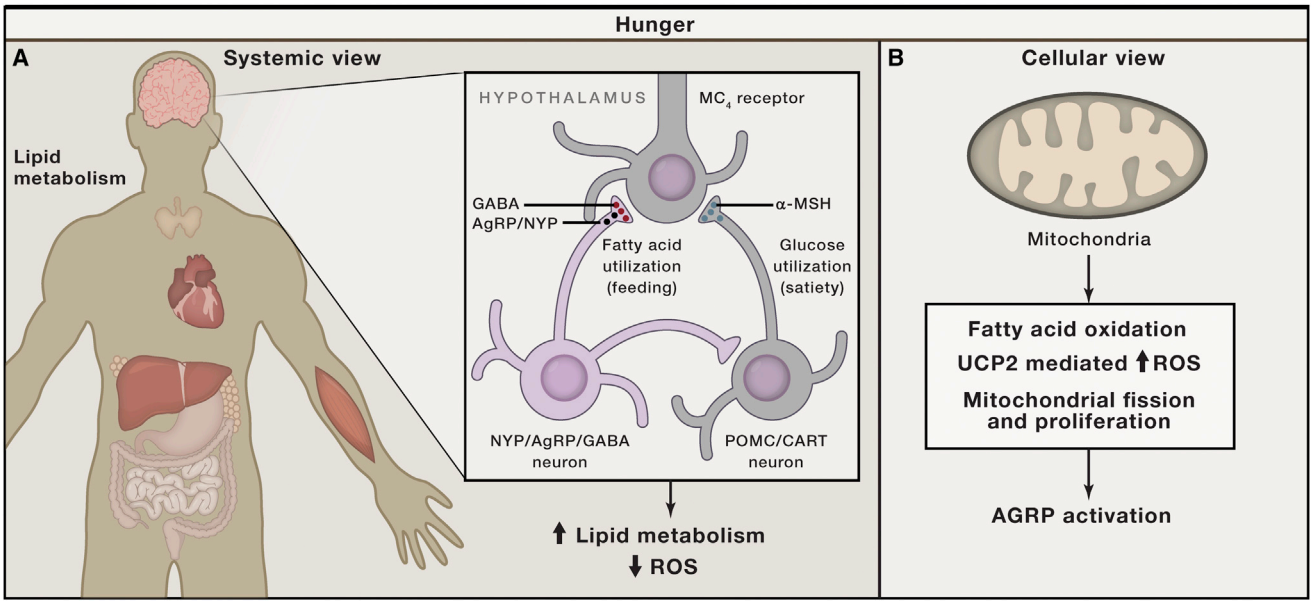


Figure 2. Schematic Illustration of Hypothalamic Control of Negative Energy Metabolism with Low ROS

(A) In the brain, the hypothalamus contains neuronal populations that control hunger (negative energy balance) and satiety (positive energy balance). Hunger state is promoted by neurons (purple) that produce Agouti-related peptide (AgRP) and neuropeptide Y (NPY), as well as GABA. When these neurons are active (hunger, calorie restriction, starvation), systemic metabolism is shifting to lipid metabolism with an overall lower level of mtROS production in all tissues.

(B) The activation of AgRP neurons during negative energy balance is promoted by pathways enabling long-chain fatty acid oxidation in the mitochondria, which is enabled by maintenance of low mtROS generation by engagement of UCP2 and mechanisms that propagate fission and/or proliferation of mitochondria (NRF1, Sirt1, and PGC1α).

behavior, energy expenditure, and glucose homeostasis (Andrews et al., 2008; Benani et al., 2007; Diano et al., 2011; Horvath et al., 2009; Leloup et al., 2006; Long et al., 2014).

Modulation of mtROS Production by Uncoupling Protein 2 in the Brain

The discovery of new members of the uncoupling protein (UCP) family in 1997 (Fleury et al., 1997) and the localization of UCP2 to specific brain areas (Horvath et al., 1999; Richard et al., 1998) initiated studies by several groups to unmask what these proteins might do in neurons and in other brain cells. Regardless of the wealth of information gained in these studies, there remains a great ambiguity about the precise role of these UCPs in cellular functions (Brand and Esteves, 2005).

There is little debate regarding the functional relevance of UCP1 in brown adipose tissue, where it promotes mitochondrial fatty acid oxidation by uncoupling mitochondrial electron transport from ATP generation under adrenergic and thyroid control (Ricquier, 1998). Under this scenario, the gained energy is dissipated in the form of heat, which is the hallmark of non-shivering thermogenesis (Ricquier, 1998). UCP2 is clearly not a classical uncoupler like UCP1, and its precise mode of action remains unclear. Furthermore, none of the cells in the brain that express UCP2 displays the aforementioned features of brown adipocytes. At baseline, UCP2 in rodents and primates is expressed predominantly in neurons of basal structures of the brain (Diano et al., 2000; Horvath et al., 1999; Richard et al., 1998). However, UCP2 is induced in many brain sites in response to cellular injury inflicted by physical insults (Bechmann et al., 2002), epileptic seizures (Diano et al., 2003), or ischemia (Deierborg et al., 2008). These seminal observations, together with studies unmasking the bidirectional regulatory relationship between UCP2 and ROS (Arsenijevic et al., 2000; Echtay et al., 2002; Negre-Salvayre et al., 1997), underscore the potential importance of this mitochondrial protein in mtROS control during cellular stress. Pursuit of the physiological functions of UCP2 in normal brain gave further support for this notion.

Initial studies in normal brain implicated UCP2 protein expression in specific subpopulations of neurons that control hunger, energy expenditure, glucose metabolism, and circadian rhythms (Horvath et al., 1999). Simultaneously, but independently, the sites of action of the hunger-promoting peripheral hormone, ghrelin (Cowley et al., 2003), revealed virtually complete overlap of ghrelin action in the brain and UCP2 expression. Eventually it became clear that most cells in the brain that express ghrelin receptors also express UCP2 (Andrews et al., 2008). These observations spurred the interrogation of whether the influence of ghrelin on appetite and food intake is mediated by UCP2 and, if so, via what cellular and intercellular mechanisms (Andrews et al., 2008; Diano and Horvath, 2012; Horvath et al., 2009). Through these investigations, the following chain of events was uncovered regarding UCP2, mtROS, and neuronal activity (Figure 2): (1) ghrelin induces NPY/AgRP neuronal firing via activation of its receptor, GHSR (growth hormone secretogouge receptor), which in turn activates AMP kinase (AMPK); (2) AMPK activation suppresses acetyl CoA carboxylase (ACC) activity, eliminating the inhibitory effect of malonyl-CoA on carnitine palmitoyl transferase 1 (CPT1) activity; (3) CPT1 activation

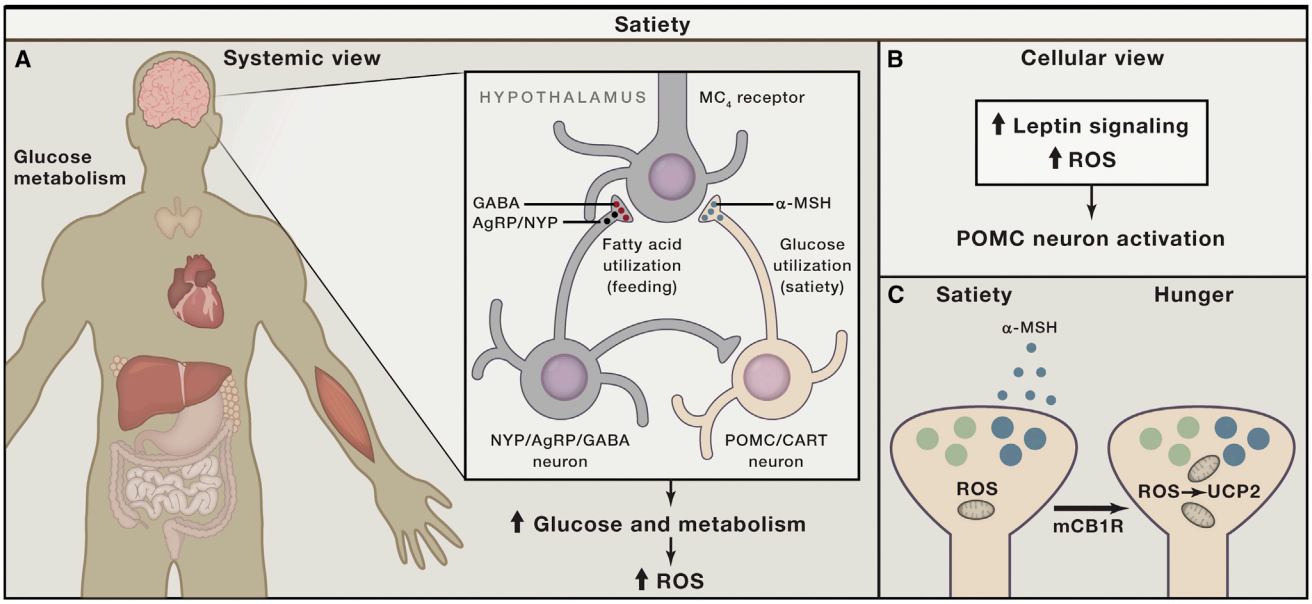


Figure 3. Schematic Illustration of Hypothalamic Control of Positive Energy Metabolism with Elevated ROS

(A) Satiety (feeling full) is promoted by hypothalamic neurons (beige) that produce pro-opiomelanocortin (POMC)-derived peptides, such as α-MSH, which, in turn, act on melanocortin-4-receptor-containing neurons. When these neurons are active, systemic metabolism is shifting toward glucose utilization, with enhanced mtROS production contributing to increased cellular ROS in various tissues.

(B) The activation of POMC neurons after a meal is accomplished by ROS, in part driven by increased mtROS production, and is supported by intracellular leptin (Jak/Stat) and insulin (PI-3K/PTEN) signaling, involving altered K-ATP channel activity.

(C) Recent studies indicate that, under unique circumstances (e.g., activation of cannabinoid receptors; mCB1R), POMC neurons, while still driven by ROS, will become promoters of hunger rather than satiety because they shift to releasing appetite-stimulating opiates (β-endorphin) though UCP2-dependent mitochondrial adaptations.

enhances long-chain fatty acid oxidation by mitochondria and the generation of mtROS; (4) ROS, together with fatty acids, promotes UCP2 gene transcription and activity; (5) UCP2, via enhancing proton leak, tempers mtROS production, allowing continuous fatty acid oxidation without oxidative stress burden and transcription of genes that promote mitochondrial biogenesis and activity (e.g., NRF1), enabling continuous support of the bioenergetic needs of sustained firing of NPY/AgRP cells; and (6) activity of NPY/AgRP neurons results in activity-dependent synaptic plasticity and inhibition of POMC neurons. The intracellular signaling of AgRP neurons is supportive of neuronal activation and decreases vulnerability of these neurons to cellular stress. It was also suggested that the long-chain fatty acyl CoAs utilized under these conditions arise from the periphery under the control of the hypothalamic NPY/AgRP neurons (Andrews et al., 2008). Knocking out UCP2 diminished the ability of AgRP neurons to inhibit mtROS production and maintain low levels of cellular ROS, which, in turn, impairs their neuronal functions (Andrews et al., 2008). Consistent with this model, cellular and electric activity of AgRP neurons is restored, together with reversal of impaired feeding behavior, when a ROS-scavenging cocktail containing L-cysteine is infused into the parenchyma of the hypothalamus (Andrews et al., 2008). These results strongly indicate that behaviors associated with low-energy availability hinge significantly on mtROS signaling associated with lipid metabolism in key neurons that drive these organismal adaptations. However, we recognize that UCP2 likely controls mtROS indirectly, via alteration of mitochondrial fuel utilization, and that other consequences of UCP2 activity in regulating metabolism may also be important (Andrews et al., 2008; Diano and Horvath, 2012; Horvath et al., 2009; Pecqueur et al., 2009; Vozza et al., 2014).

The studies outlined above also suggested the exact opposite scenario for those neurons that support cessation of eating once enough food is consumed (satiety). The hypothalamic neurons that produce pro-opiomelanocortin (POMC) and related peptides are located in the same area as the hunger-promoting AgRP neurons that, when active, suppress POMC neuronal activity. POMC neurons appear to have elevated ROS, as indicated by intracellular dihydroethidium (DHE) staining (Andrews et al., 2008; Diano et al., 2011), when they are active to promote satiety and increased energy expenditure after food consumption (Figure 3). Elevated mtROS production is a logical POMC neuronal activator, since it would likely correlate with mitochondrial activation during full oxidation of glucose, the main fuel of these neurons (Parton et al., 2007). When ROS levels are suppressed chemically, POMC neurons are hyperpolarized and their firing rate declines (Diano et al., 2011). Conversely, when in-slice preparations of POMC neurons are exposed to hydrogen peroxide, they become depolarized and their firing rate is elevated (Diano et al., 2011). These results indicate that it is actually mtROS, rather than glucose itself, that instigate POMC neuronal firing (Diano et al., 2011; Long et al., 2014). Because hypothalamic POMC neurons are involved in both behavioral and

autonomic control of energy and glucose metabolism, it is not surprising that hypothalamic ROS control has been tied to all of these processes (Figure 3).

ROS are Satiety Signals

From a behavioral and systemic perspective, under normal conditions, elevated hypothalamic ROS levels are permissive for suppression of eating, increased energy expenditure, and glucose utilization by peripheral tissues (Andrews et al., 2008; Benani et al., 2007; Diano et al., 2011; Leloup et al., 2006; Long et al., 2014). In fact, it is reasonable to conclude that ROS signaling in the brain, as well as in the periphery, is fundamental for appropriate behavioral and autonomic adaptations to energy surplus (e.g., after consumption of a meal). If one interferes with ROS in these physiological processes, both behavioral and autonomic correlates of proper fuel management of the body will be impaired (Andrews et al., 2008; Diano et al., 2011; Horvath et al., 2009; Long et al., 2014). Since mitochondrial perturbations via UCP2 modulate these responses, we conclude that mtROS signaling plays a fundamental and crucial role in physiological regulation of systemic metabolism.

The relentless pursuit of available energy sources in the environment is mandatory for organismal survival, and hence, hunger and hunger-controlled circuits motivate this critical behavior. However, advanced animal species, including humans, also developed the capacity to maintain energy reserves, for example, in the form of fat or glycogen, so they do not have to continuously feed to survive. This evolutionary advantage, by default, demands that ''gauges'' and ''switches'' are in place to shut off feeding when storage capacity is sufficient. The role of the hypothalamic melanocortin system appears to represent both the gauge and the switch, with ROS, likely driven by mtROS production, being the sensor. When cellular ROS reach a threshold, they activate POMC neurons, enabling the cessation of feeding and initiation of storage and utilization of fuels via processes controlled by insulin and leptin (Varela and Horvath, 2012). The control of insulin release itself is regulated by redox events (Bashan et al., 2009). At the same time, elevating ROS in the hypothalamus reduces activity of AgRP neurons that propagate hunger (Andrews et al., 2008) (Figure 2). While it remains unknown what underlies the differential effect of ROS on POMC and AgRP neurons, cell-specific expression of plasma membrane channels could be involved. For example, superoxide can directly alter the activity of potassium channels (Avshalumov and Rice, 2003), which play important roles in glucose sensing by hypothalamic neurons and under regulatory control by UCP2 (Parton et al., 2007). Likewise, it remains unclear how mtROS signals reach neuronal perikarya. In this regard, it is noteworthy that mitochondria in both AgRP and POMC neurons dynamically change their morphology and localization within a short period of time (Andrews et al., 2008; Coppola et al., 2007; Dietrich et al., 2013; Schneeberger et al., 2013). Mitochondrial fission and fusion capacity and the physical interaction between mitochondria and the endoplasmic reticulum may be involved in mtROS production or associated signaling events (Dietrich et al., 2013; Nasrallah and Horvath, 2014; Schneeberger et al., 2013).

It is important to note that ROS are likely not the satiety signal in the hypothalamus under all circumstances. For example, the known effect of cannabinoids on promoting ferocious appetite, regardless of metabolic satiety, is actually mediated by mtROS-driven POMC neurons (Koch et al., 2015). However, under cannabinoid influence, POMC neurons reverse their function and promote appetite because they switch from the release of its satiety-promoting neuropeptide, α-melanocyte-stimulating hormone (α-MSH), to β-endorphin (Koch et al., 2015). This switch is enabled by a mitochondrial adaptive response controlled by UCP2 (Koch et al., 2015) (Figure 3C). Whether this response is specific for this pharmacological situation or relevant to regulation of metabolism under certain physiological circumstances is unknown.

Mitochondrial ROS and Exercise

A key feature of animal species is their need to physically relocate in a rapid and predictable fashion to find food, reproduce, or escape danger. This is accomplished, in grosso modo, through behavioral adaptations, which are the sum of coordinating sensory and effector functions via the communication between the nervous system (the sensory component) and the musculoskeletal system (the effector component). Movement and exercise, in general, have long been recognized as key to supporting not only the aforementioned fundamental biological needs, but also tissue health and longevity. Ristow and colleagues showed that suppressing ROS generation during aerobic exercise (likely, in large part, mtROS) diminishes beneficial outcomes on many exercise-related parameters (Ristow et al., 2009). They also argue for the critical relevance of mtROS signaling transients as important contributors to longevity, as well as mediators of other signaling responses that promote healthspan and longevity (Schmeisser et al., 2013; Zarse et al., 2012). In support of the notion that mtROS transients (and not sustained elevated ROS levels) mediate the benefits of exercise on integrative physiology, UCP2 was found crucial to support exercise-induced synaptogenesis in the dentate gyrus of the hippocampal formation, a key site of spatial learning (Dietrich et al., 2008). This same mechanism is also relevant to hippocampal development (Simon-Areces et al., 2012) and lifespan determination (Andrews and Horvath, 2009).

Short- and Long-Term Effects of ROS

The distinction between physiologically beneficial short ROS bursts and pathological, sustained high ROS levels on cellular and circuit integrity is best illustrated in the response of the hypothalamus to exposure to calorie-dense diets containing high levels of fats and carbohydrates (Diano et al., 2011; Parton et al., 2007). On regular chow diet, which contains <20% fat, ROS levels fluctuate between hunger and satiety states and mice maintain a positive correlation between hypothalamic ROS levels, circulating leptin levels (the adipose hormone that signals to the hypothalamus when sufficient amount of food is consumed), and activity of POMC neurons (Diano et al., 2011). However, when animals are placed on calorie-dense diets (>40% fat), homeostatic control of energy metabolism is gradually deregulated. When this occurs, animals steadily increase their fat stores, which results in steady elevation of circulating

leptin. Under homeostatic conditions, this elevated leptin would decrease feeding and enable maintenance of fat stores. This inability of elevated leptin levels to avoid or reverse weight gain is called leptin resistance, for which many cellular and tissue mechanisms have been proposed to explain. One of these mechanisms relates to the aforementioned mtROS control in POMC and AgRP neurons. On a high-fat diet, the positive correlation between hypothalamic ROS, circulating leptin levels, and POMC neuronal activity in mice deteriorates (Diano et al., 2011). Under these conditions, hypothalamic ROS levels plateau and do not follow the robust and steady elevation in circulating leptin concentrations. At the same time, POMC neuronal activity is diminished (Diano et al., 2011). The underlying cause for this dysregulation is tied to PPARγ-related proliferation of peroxisomes and increased ROS elimination by catalase (Diano et al., 2011; Long et al., 2014). We suggest that leptin affects this process by enabling increased glucose uptake in POMC and AgRP neurons, which is accompanied by increased lipid load. Initially, this will lead to increased mtROS production and crossing cellular ROS thresholds that promote satiety by activating POMC neurons. However, rising leptin levels on high-fat diet will continue to promote glucose uptake by these neurons, which, together with increasing lipid load, will overburden these postmitotic cells. The scenario in which lipid and carbohydrate load is increasing in cells provides a perfect energetic basis for growth. Carbohydrate oxidation will dominate mitochondrial OXPHOS and ATP generation, while long-chain fatty acids are diverted from mitochondria via the malonyl-CoA shuttle for biogenesis of membranes and cell growth. However, in cells whose growth is strictly limited, such as neurons of the adult central nervous system, lipids cannot be continuously utilized for membrane biogenesis and they will accumulate in various intracellular compartments, including the endoplasmic reticulum. The activation of peroxisome proliferation under these circumstances provides an alternative mechanism through which excess fat within cells can be eliminated via mitochondrial β oxidation that is less coupled to ATP generation. While this is a beneficial process to prevent lipotoxicity, peroxisomal catalase activity will limit ROS generation needed for proper signaling from the hypothalamus to diminish feeding and increase energy expenditure (Diano et al., 2011). These alterations may have multiple negative effects on cellular signaling mechanisms and organelle integrity and function.

A "Fuel Hypothesis" of Cellular Function

While the above details on mtROS-related mechanisms were described in relation to a hypothalamic circuit that controls feeding behavior and peripheral fuel partitioning and utilization, these processes are likely relevant to the functionality and impairment of neurons in various parts of the brain. For example, UCP2-dependent control of mtROS was also found in dopamine neurons in the midbrain substantia nigra, where both normal functioning of these cells and their protection under cellular stress were attributed to this mechanism (Andrews et al., 2005, 2006; Conti et al., 2005). Dopamine cells in this area of the brain are connected to control of fine motor functions and complex motivated behaviors. A role for the peripheral metabolic hormone, ghrelin, was identified to modulate the activity of these neurons and to prevent their impairment and death in models of Parkinson's disease (Abizaid et al., 2006; Andrews et al., 2009). The intracellular signaling pathway that enabled these beneficial effects of ghrelin was related to ROS control by the same machinery as described above in relation to the control of feeding (Andrews et al., 2008). Similar ghrelin action was also found in the hippocampus to promote learning and memory and to ameliorate deficits of animals in a model of Alzheimer's disease (Diano et al., 2006). Because the changing metabolic state (hunger ↔ satiety) is closely tied to predictable changes in complex behaviors, it is reasonable to suggest that fluctuating ROS levels (mediated by alterations in mtROS output) in all or part of the brain play a critical regulatory role in the synchronization of neuronal circuit activity in support of continuous and appropriate behavioral adaptations. Furthermore, mtROS-controlled neuronal activity in the hypothalamus is sufficient to affect complex behaviors beyond feeding. For example, actute activation of hypothalamic AgRP neurons rapidly alters stereotypic behaviors, locomotion, and anxiety (Dietrich et al., 2015). Finally, we speculate that purposeful alterations in mtROS production to effect cellular redox signaling pathways (Figure 1) will regulate homeostasis in other tissues. In simple terms, it is reasonable to assume that cellular functions in any tissue are determined by fuel availability, uptake, and utilization and that these "fuel" principles drive and orchestrate signaling modalities, including mtROS signaling, to control homeostatic and adaptive responses. For example, intracellular metabolic pathways have distinct and dominant impacts on various immune cell types (Caro-Maldonado et al., 2012; Procaccini et al., 2010), and UCP2-dependent mtROS regulation is connected to both adaptive and innate immune cell functions (Arsenijevic et al., 2000; Horvath et al., 2003; Krauss et al., 2002). How other cells and tissues respond to such signals and the intersection between cell-intrinsic signaling and control by the CNS is an exciting area of future research. In this regard, determining whether the aforementioned CNS processes are mediated by cell-non-autonomous mtROS-mediated signals similar to those documented in *C. elegans* downstream of ETC disruption and mtROS (Durieux et al., 2011; Schieber and Chandel, 2014) will be important to consider.

Challenges in ROS Research and Ramifications of ROS as Central Controllers of Organismal Homeostasis

In this Review, we have summarized how changes in mtROS production can impact cellular ROS thresholds and redox signaling events that control basal physiological functions and adaptive responses. As such, we argue that these pathways are critical for organismal homeostasis, stress responsiveness, health, and longevity. However, these pathways are far from understood and are in need of more intensive study. Some current impediments and other relevant considerations as the field moves forward in this area are covered below.

At present, it remains very difficult to effectively measure ROS in cells and in vivo. The use of commercially fluorescent ROS probes is widespread but often without the knowledge that these do not always readout specific ROS species faithfully and are prone to other confounding artifacts (Kalyanaraman et al., 2012). That is not to say that these are not useful to a degree,

but they should not be used as the only line of evidence to implicate ROS in a response. Development of better ROS assays and probes is ongoing (Ezeriņa et al., 2014; Logan et al., 2014; Woolley et al., 2013), yet there remains a great need for additional forward progress in this important area.

There are important implications for ROS as physiological signaling molecules that impact therapeutic strategies. Two that immediately come to mind are the use of antioxidants and anti-obesity strategies that target the CNS. Antioxidants have long been considered potential therapeutics for a number of conditions involving oxidative stress. In general, trials using these have failed, likely, in part, because of unintentional inhibition of important basal and adaptive ROS signaling pathways. It has even been argued that taking antioxidants as daily dietary supplements might also perturb these ROS pathways in ways that are not beneficial or even harmful (Ristow, 2014). Similarly, strategies that target activation of POMC neurons (to promote satiety) as an anti-obesity/anti-diabetic strategy might also be confounded due to perturbation of ROS signaling circuits that we have described herein (Dietrich and Horvath, 2012). That is, we assert that many compounds that activate POMC neurons will, by default, upregulate ROS production and signaling that governs their activity (Diano et al., 2011). If this scenario is maintained for a prolonged period of time (hours, days, months), weight loss may be accomplished, but sustained ROS levels could have a multitude of unintended detrimental consequences. As these examples point out, previously held views of ROS as just damaging agents that need to be eliminated are out of date, and a new appreciation of their signaling roles is important to consider going forward. The fact that mitochondria are major producers of ROS also highlights the importance of better understanding what controls their activity and rate of mtROS production, both in terms of redox signaling and oxidative stress. In this regard, the greater recent appreciation of mitochondria as important signaling hubs (Chandel, 2014; West et al., 2011) has begun to transcend older, over-simplified views of these organelles as just sites of intermediary metabolism and ATP production.

REFERENCES

Abizaid, A., Liu, Z.W., Andrews, Z.B., Shanabrough, M., Borok, E., Elsworth, J.D., Roth, R.H., Sleeman, M.W., Picciotto, M.R., Tschöp, M.H., et al. (2006). Ghrelin modulates the activity and synaptic input organization of midbrain dopamine neurons while promoting appetite. J. Clin. Invest. *116*, 3229–3239.

Al-Mehdi, A.B., Pastukh, V.M., Swiger, B.M., Reed, D.J., Patel, M.R., Bardwell, G.C., Pastukh, V.V., Alexeyev, M.F., and Gillespie, M.N. (2012). Perinuclear mitochondrial clustering creates an oxidant-rich nuclear domain required for hypoxia-induced transcription. Sci. Signal. *5*, ra47.

Andrews, Z.B., Erion, D., Beiler, R., Liu, Z.W., Abizaid, A., Zigman, J., Elsworth, J.D., Savitt, J.M., DiMarchi, R., Tschoep, M., et al. (2009). Ghrelin promotes and protects nigrostriatal dopamine function via a UCP2-dependent mitochondrial mechanism. J. Neurosci. *29*, 14057–14065.

Andrews, Z.B., Horvath, B., Barnstable, C.J., Elsworth, J., Yang, L., Beal, M.F., Roth, R.H., Matthews, R.T., and Horvath, T.L. (2005). Uncoupling protein-2 is critical for nigral dopamine cell survival in a mouse model of Parkinson's disease. J. Neurosci. *25*, 184–191.

Andrews, Z.B., and Horvath, T.L. (2009). Uncoupling protein-2 regulates lifespan in mice. Am. J. Physiol. Endocrinol. Metab. *296*, E621–E627.

Andrews, Z.B., Liu, Z.W., Walllingford, N., Erion, D.M., Borok, E., Friedman, J.M., Tschöp, M.H., Shanabrough, M., Cline, G., Shulman, G.I., et al. (2008). UCP2 mediates ghrelin's action on NPY/AgRP neurons by lowering free radicals. Nature *454*, 846–851.

Andrews, Z.B., Rivera, A., Elsworth, J.D., Roth, R.H., Agnati, L., Gago, B., Abizaid, A., Schwartz, M., Fuxe, K., and Horvath, T.L. (2006). Uncoupling protein-2 promotes nigrostriatal dopamine neuronal function. Eur. J. Neurosci. *24*, 32–36.

Arsenijevic, D., Onuma, H., Pecqueur, C., Raimbault, S., Manning, B.S., Miroux, B., Couplan, E., Alves-Guerra, M.C., Goubern, M., Surwit, R., et al. (2000). Disruption of the uncoupling protein-2 gene in mice reveals a role in immunity and reactive oxygen species production. Nat. Genet. *26*, 435–439.

Avshalumov, M.V., and Rice, M.E. (2003). Activation of ATP-sensitive K+ (K(ATP)) channels by H2O2 underlies glutamate-dependent inhibition of striatal dopamine release. Proc. Natl. Acad. Sci. USA *100*, 11729–11734.

Bashan, N., Kovsan, J., Kachko, I., Ovadia, H., and Rudich, A. (2009). Positive and negative regulation of insulin signaling by reactive oxygen and nitrogen species. Physiol. Rev. *89*, 27–71.

Bechmann, I., Diano, S., Warden, C.H., Bartfai, T., Nitsch, R., and Horvath, T.L. (2002). Brain mitochondrial uncoupling protein 2 (UCP2): a protective stress signal in neuronal injury. Biochem. Pharmacol. *64*, 363–367.

Benani, A., Troy, S., Carmona, M.C., Fioramonti, X., Lorsignol, A., Leloup, C., Casteilla, L., and Pénicaud, L. (2007). Role for mitochondrial reactive oxygen species in brain lipid sensing: redox regulation of food intake. Diabetes *56*, 152–160.

Bonawitz, N.D., Chatenay-Lapointe, M., Pan, Y., and Shadel, G.S. (2007). Reduced TOR signaling extends chronological life span via increased respiration and upregulation of mitochondrial gene expression. Cell Metab. *5*, 265–277.

Brand, M.D., and Esteves, T.C. (2005). Physiological functions of the mitochondrial uncoupling proteins UCP2 and UCP3. Cell Metab. *2*, 85–93.

Budzińska, M., Gałgańska, H., Karachitos, A., Wojtkowska, M., and Kmita, H. (2009). The TOM complex is involved in the release of superoxide anion from mitochondria. J. Bioenerg. Biomembr. *41*, 361–367.

Butow, R.A., and Avadhani, N.G. (2004). Mitochondrial signaling: the retrograde response. Mol. Cell *14*, 1–15.

Calvo, S.E., and Mootha, V.K. (2010). The mitochondrial proteome and human disease. Annu. Rev. Genomics Hum. Genet. *11*, 25–44.

Caro-Maldonado, A., Gerriets, V.A., and Rathmell, J.C. (2012). Matched and mismatched metabolic fuels in lymphocyte function. Semin. Immunol. *24*, 405–413.

Chandel, N.S. (2014). Mitochondria as signaling organelles. BMC Biol. *12*, 34.

Conti, B., Sugama, S., Lucero, J., Winsky-Sommerer, R., Wirz, S.A., Maher, P., Andrews, Z., Barr, A.M., Morale, M.C., Paneda, C., et al. (2005). Uncoupling protein 2 protects dopaminergic neurons from acute 1,2,3,6-methyl-phenyl-tetrahydropyridine toxicity. J. Neurochem. *93*, 493–501.

Coppola, A., Liu, Z.W., Andrews, Z.B., Paradis, E., Roy, M.C., Friedman, J.M., Ricquier, D., Richard, D., Horvath, T.L., Gao, X.B., and Diano, S. (2007). A central thermogenic-like mechanism in feeding regulation: an interplay between arcuate nucleus T3 and UCP2. Cell Metab. *5*, 21–33.

Cowley, M.A., Smith, R.G., Diano, S., Tschöp, M., Pronchuk, N., Grove, K.L., Strasburger, C.J., Bidlingmaier, M., Esterman, M., Heiman, M.L., et al. (2003). The distribution and mechanism of action of ghrelin in the CNS demonstrates a novel hypothalamic circuit regulating energy homeostasis. Neuron *37*, 649–661.

Dancy, B.M., Sedensky, M.M., and Morgan, P.G. (2014). Effects of the mitochondrial respiratory chain on longevity in C. elegans. Exp. Gerontol. *56*, 245–255.

Deierborg, T., Wieloch, T., Diano, S., Warden, C.H., Horvath, T.L., and Mattiasson, G. (2008). Overexpression of UCP2 protects thalamic neurons following global ischemia in the mouse. Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism *28*, 1186–1195.

Diano, S., and Horvath, T.L. (2012). Mitochondrial uncoupling protein 2 (UCP2) in glucose and lipid metabolism. Trends Mol. Med. *18*, 52–58.

Diano, S., Urbanski, H.F., Horvath, B., Bechmann, I., Kagiya, A., Nemeth, G., Naftolin, F., Warden, C.H., and Horvath, T.L. (2000). Mitochondrial uncoupling protein 2 (UCP2) in the nonhuman primate brain and pituitary. Endocrinology *141*, 4226–4238.

Diano, S., Matthews, R.T., Patrylo, P., Yang, L., Beal, M.F., Barnstable, C.J., and Horvath, T.L. (2003). Uncoupling protein 2 prevents neuronal death including that occurring during seizures: a mechanism for preconditioning. Endocrinology *144*, 5014–5021.

Diano, S., Farr, S.A., Benoit, S.C., McNay, E.C., da Silva, I., Horvath, B., Gaskin, F.S., Nonaka, N., Jaeger, L.B., Banks, W.A., et al. (2006). Ghrelin controls hippocampal spine synapse density and memory performance. Nat. Neurosci. *9*, 381–388.

Diano, S., Liu, Z.W., Jeong, J.K., Dietrich, M.O., Ruan, H.B., Kim, E., Suyama, S., Kelly, K., Gyengesi, E., Arbiser, J.L., et al. (2011). Peroxisome proliferation-associated control of reactive oxygen species sets melanocortin tone and feeding in diet-induced obesity. Nat. Med. *17*, 1121–1127.

Dietrich, M.O., Andrews, Z.B., and Horvath, T.L. (2008). Exercise-induced synaptogenesis in the hippocampus is dependent on UCP2-regulated mitochondrial adaptation. J. Neurosci. *28*, 10766–10771.

Dietrich, M.O., and Horvath, T.L. (2012). Limitations in anti-obesity drug development: the critical role of hunger-promoting neurons. Nat. Rev. Drug Discov. *11*, 675–691.

Dietrich, M.O., Liu, Z.W., and Horvath, T.L. (2013). Mitochondrial dynamics controlled by mitofusins regulate Agrp neuronal activity and diet-induced obesity. Cell *155*, 188–199.

Dietrich, M.O., Zimmer, M.R., Bober, J., and Horvath, T.L. (2015). Hypothalamic Agrp neurons drive stereotypic behaviors beyond feeding. Cell *160*, 1222–1232.

Durieux, J., Wolff, S., and Dillin, A. (2011). The cell-non-autonomous nature of electron transport chain-mediated longevity. Cell *144*, 79–91.

Echtay, K.S., Murphy, M.P., Smith, R.A., Talbot, D.A., and Brand, M.D. (2002). Superoxide activates mitochondrial uncoupling protein 2 from the matrix side. Studies using targeted antioxidants. J. Biol. Chem. *277*, 47129–47135.

Ezeriņa, D., Morgan, B., and Dick, T.P. (2014). Imaging dynamic redox processes with genetically encoded probes. J. Mol. Cell. Cardiol. *73*, 43–49.

Finkel, T. (2012). Signal transduction by mitochondrial oxidants. J. Biol. Chem. *287*, 4434–4440.

Fleury, C., Neverova, M., Collins, S., Raimbault, S., Champigny, O., Levi-Meyrueis, C., Bouillaud, F., Seldin, M.F., Surwit, R.S., Ricquier, D., and Warden, C.H. (1997). Uncoupling protein-2: a novel gene linked to obesity and hyperinsulinemia. Nat. Genet. *15*, 269–272.

Go, Y.M., Chandler, J.D., and Jones, D.P. (2015). The cysteine proteome. Free Radic. Biol. Med. *84*, 227–245.

Hamanaka, R.B., and Chandel, N.S. (2010). Mitochondrial reactive oxygen species regulate cellular signaling and dictate biological outcomes. Trends Biochem. Sci. *35*, 505–513.

Hamanaka, R.B., Glasauer, A., Hoover, P., Yang, S., Blatt, H., Mullen, A.R., Getsios, S., Gottardi, C.J., DeBerardinis, R.J., Lavker, R.M., and Chandel, N.S. (2013). Mitochondrial reactive oxygen species promote epidermal differentiation and hair follicle development. Sci. Signal. *6*, ra8.

Han, D., Antunes, F., Canali, R., Rettori, D., and Cadenas, E. (2003). Voltage-dependent anion channels control the release of the superoxide anion from mitochondria to cytosol. J. Biol. Chem. *278*, 5557–5563.

Haynes, C.M., Fiorese, C.J., and Lin, Y.F. (2013). Evaluating and responding to mitochondrial dysfunction: the mitochondrial unfolded-protein response and beyond. Trends Cell Biol. *23*, 311–318.

Hekimi, S., Lapointe, J., and Wen, Y. (2011). Taking a "good" look at free radicals in the aging process. Trends Cell Biol. *21*, 569–576.

Horvath, T.L., Andrews, Z.B., and Diano, S. (2009). Fuel utilization by hypothalamic neurons: roles for ROS. Trends Endocrinol. Metab. *20*, 78–87.

Horvath, T.L., Diano, S., Miyamoto, S., Barry, S., Gatti, S., Alberati, D., Livak, F., Lombardi, A., Moreno, M., Goglia, F., et al. (2003). Uncoupling proteins-2 and 3 influence obesity and inflammation in transgenic mice. International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity *27*, 433–442.

Horvath, T.L., Warden, C.H., Hajos, M., Lombardi, A., Goglia, F., and Diano, S. (1999). Brain uncoupling protein 2: uncoupled neuronal mitochondria predict thermal synapses in homeostatic centers. J. Neurosci. *19*, 10417–10427.

Hwang, A.B., Ryu, E.A., Artan, M., Chang, H.W., Kabir, M.H., Nam, H.J., Lee, D., Yang, J.S., Kim, S., Mair, W.B., et al. (2014). Feedback regulation via AMPK and HIF-1 mediates ROS-dependent longevity in Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA *111*, E4458–E4467.

Johnson, D., Allman, E., and Nehrke, K. (2012). Regulation of acid-base transporters by reactive oxygen species following mitochondrial fragmentation. Am. J. Physiol. Cell Physiol. *302*, C1045–C1054.

Kalyanaraman, B., Darley-Usmar, V., Davies, K.J., Dennery, P.A., Forman, H.J., Grisham, M.B., Mann, G.E., Moore, K., Roberts, L.J., 2nd, and Ischiropoulos, H. (2012). Measuring reactive oxygen and nitrogen species with fluorescent probes: challenges and limitations. Free Radic. Biol. Med. *52*, 1–6.

Koch, M., Varela, L., Kim, J.G., Kim, J.D., Hernández-Nuño, F., Simonds, S.E., Castorena, C.M., Vianna, C.R., Elmquist, J.K., Morozov, Y.M., et al. (2015). Hypothalamic POMC neurons promote cannabinoid-induced feeding. Nature *519*, 45–50.

Krauss, S., Zhang, C.Y., and Lowell, B.B. (2002). A significant portion of mitochondrial proton leak in intact thymocytes depends on expression of UCP2. Proc. Natl. Acad. Sci. USA *99*, 118–122.

Labbé, K., Murley, A., and Nunnari, J. (2014). Determinants and functions of mitochondrial behavior. Annu. Rev. Cell Dev. Biol. *30*, 357–391.

Lee, S.J., Hwang, A.B., and Kenyon, C. (2010). Inhibition of respiration extends C. elegans life span via reactive oxygen species that increase HIF-1 activity. Curr. Biol. *20*, 2131–2136.

Leloup, C., Magnan, C., Benani, A., Bonnet, E., Alquier, T., Offer, G., Carriere, A., Périquet, A., Fernandez, Y., Ktorza, A., et al. (2006). Mitochondrial reactive oxygen species are required for hypothalamic glucose sensing. Diabetes *55*, 2084–2090.

Logan, A., Cochemé, H.M., Li Pun, P.B., Apostolova, N., Smith, R.A., Larsen, L., Larsen, D.S., James, A.M., Fearnley, I.M., Rogatti, S., et al. (2014). Using exomarkers to assess mitochondrial reactive species in vivo. Biochim. Biophys. Acta *1840*, 923–930.

Long, L., Toda, C., Jeong, J.K., Horvath, T.L., and Diano, S. (2014). PPARγ ablation sensitizes proopiomelanocortin neurons to leptin during high-fat feeding. J. Clin. Invest. *124*, 4017–4027.

Lustgarten, M.S., Bhattacharya, A., Muller, F.L., Jang, Y.C., Shimizu, T., Shirasawa, T., Richardson, A., and Van Remmen, H. (2012). Complex I generated, mitochondrial matrix-directed superoxide is released from the mitochondria through voltage dependent anion channels. Biochem. Biophys. Res. Commun. *422*, 515–521.

Mishra, P., and Chan, D.C. (2014). Mitochondrial dynamics and inheritance during cell division, development and disease. Nat. Rev. Mol. Cell Biol. *15*, 634–646.

Murphy, M.P. (2009). How mitochondria produce reactive oxygen species. Biochem. J. *417*, 1–13.

Nasrallah, C.M., and Horvath, T.L. (2014). Mitochondrial dynamics in the central regulation of metabolism. Nat. Rev. Endocrinol. *10*, 650–658.

Negre-Salvayre, A., Hirtz, C., Carrera, G., Cazenave, R., Troly, M., Salvayre, R., Penicaud, L., and Casteilla, L. (1997). A role for uncoupling protein-2 as a regulator of mitochondrial hydrogen peroxide generation. FASEB journal: official publication of the Federation of American Societies for Experimental Biology *11*, 809–815.

Nunnari, J., and Suomalainen, A. (2012). Mitochondria: in sickness and in health. Cell *148*, 1145–1159.

Pan, Y., Schroeder, E.A., Ocampo, A., Barrientos, A., and Shadel, G.S. (2011). Regulation of yeast chronological life span by TORC1 via adaptive mitochondrial ROS signaling. Cell Metab. *13*, 668–678.

Parton, L.E., Ye, C.P., Coppari, R., Enriori, P.J., Choi, B., Zhang, C.Y., Xu, C., Vianna, C.R., Balthasar, N., Lee, C.E., et al. (2007). Glucose sensing by POMC neurons regulates glucose homeostasis and is impaired in obesity. Nature *449*, 228–232.

Pecqueur, C., Alves-Guerra, C., Ricquier, D., and Bouillaud, F. (2009). UCP2, a metabolic sensor coupling glucose oxidation to mitochondrial metabolism? IUBMB Life *61*, 762–767.

Procaccini, C., De Rosa, V., Galgani, M., Abanni, L., Calì, G., Porcellini, A., Carbone, F., Fontana, S., Horvath, T.L., La Cava, A., and Matarese, G. (2010). An oscillatory switch in mTOR kinase activity sets regulatory T cell responsiveness. Immunity *33*, 929–941.

Quinlan, C.L., Perevoshchikova, I.V., Hey-Mogensen, M., Orr, A.L., and Brand, M.D. (2013). Sites of reactive oxygen species generation by mitochondria oxidizing different substrates. Redox Biol. *1*, 304–312.

Richard, D., Rivest, R., Huang, Q., Bouillaud, F., Sanchis, D., Champigny, O., and Ricquier, D. (1998). Distribution of the uncoupling protein 2 mRNA in the mouse brain. J. Comp. Neurol. *397*, 549–560.

Ricquier, D. (1998). Neonatal brown adipose tissue, UCP1 and the novel uncoupling proteins. Biochem. Soc. Trans. *26*, 120–123.

Ristow, M. (2014). Unraveling the truth about antioxidants: mitohormesis explains ROS-induced health benefits. Nat. Med. *20*, 709–711.

Ristow, M., and Schmeisser, S. (2011). Extending life span by increasing oxidative stress. Free Radic. Biol. Med. *51*, 327–336.

Ristow, M., and Zarse, K. (2010). How increased oxidative stress promotes longevity and metabolic health: The concept of mitochondrial hormesis (mitohormesis). Exp. Gerontol. *45*, 410–418.

Ristow, M., Zarse, K., Oberbach, A., Klöting, N., Birringer, M., Kiehntopf, M., Stumvoll, M., Kahn, C.R., and Blüher, M. (2009). Antioxidants prevent health-promoting effects of physical exercise in humans. Proc. Natl. Acad. Sci. USA *106*, 8665–8670.

Rugarli, E.I., and Langer, T. (2012). Mitochondrial quality control: a matter of life and death for neurons. EMBO J. *31*, 1336–1349.

Scarpulla, R.C. (2008). Transcriptional paradigms in mammalian mitochondrial biogenesis and function. Physiol. Rev. *88*, 611–638.

Scheibye-Knudsen, M., Fang, E.F., Croteau, D.L., Wilson, D.M., 3rd, and Bohr, V.A. (2015). Protecting the mitochondrial powerhouse. Trends Cell Biol. *25*, 158–170.

Schieber, M., and Chandel, N.S. (2014). TOR signaling couples oxygen sensing to lifespan in C. elegans. Cell Rep. *9*, 9–15.

Schmeisser, S., Priebe, S., Groth, M., Monajembashi, S., Hemmerich, P., Guthke, R., Platzer, M., and Ristow, M. (2013). Neuronal ROS signaling rather than AMPK/sirtuin-mediated energy sensing links dietary restriction to lifespan extension. Mol. Metab. *2*, 92–102.

Schneeberger, M., Dietrich, M.O., Sebastián, D., Imbernón, M., Castaño, C., Garcia, A., Esteban, Y., Gonzalez-Franquesa, A., Rodríguez, I.C., Bortolozzi, A., et al. (2013). Mitofusin 2 in POMC neurons connects ER stress with leptin resistance and energy imbalance. Cell *155*, 172–187.

Schroeder, E.A., Raimundo, N., and Shadel, G.S. (2013). Epigenetic silencing mediates mitochondria stress-induced longevity. Cell Metab. *17*, 954–964.

Schulz, T.J., Zarse, K., Voigt, A., Urban, N., Birringer, M., and Ristow, M. (2007). Glucose restriction extends Caenorhabditis elegans life span by inducing mitochondrial respiration and increasing oxidative stress. Cell Metab. *6*, 280–293.

Sena, L.A., and Chandel, N.S. (2012). Physiological roles of mitochondrial reactive oxygen species. Mol. Cell *48*, 158–167.

Shadel, G.S., and Clayton, D.A. (1997). Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. *66*, 409–435.

Simon-Areces, J., Dietrich, M.O., Hermes, G., Garcia-Segura, L.M., Arevalo, M.A., and Horvath, T.L. (2012). UCP2 induced by natural birth regulates neuronal differentiation of the hippocampus and related adult behavior. PLoS ONE *7*, e42911.

Tormos, K.V., Anso, E., Hamanaka, R.B., Eisenbart, J., Joseph, J., Kalyanaraman, B., and Chandel, N.S. (2011). Mitochondrial complex III ROS regulate adipocyte differentiation. Cell Metab. *14*, 537–544.

Varela, L., and Horvath, T.L. (2012). AgRP neurons: a switch between peripheral carbohydrate and lipid utilization. EMBO J. *31*, 4252–4254.

Vozza, A., Parisi, G., De Leonardis, F., Lasorsa, F.M., Castegna, A., Amorese, D., Marmo, R., Calcagnile, V.M., Palmieri, L., Ricquier, D., et al. (2014). UCP2 transports C4 metabolites out of mitochondria, regulating glucose and glutamine oxidation. Proc. Natl. Acad. Sci. USA *111*, 960–965.

Weimer, S., Priebs, J., Kuhlow, D., Groth, M., Priebe, S., Mansfeld, J., Merry, T.L., Dubuis, S., Laube, B., Pfeiffer, A.F., et al. (2014). D-Glucosamine supplementation extends life span of nematodes and of ageing mice. Nat. Commun. *5*, 3563.

West, A.P., Khoury-Hanold, W., Staron, M., Tal, M.C., Pineda, C.M., Lang, S.M., Bestwick, M., Duguay, B.A., Raimundo, N., MacDuff, D.A., et al. (2015). Mitochondrial DNA stress primes the antiviral innate immune response. Nature *520*, 553–557.

West, A.P., Shadel, G.S., and Ghosh, S. (2011). Mitochondria in innate immune responses. Nat. Rev. Immunol. *11*, 389–402.

Winterbourn, C.C. (2008). Reconciling the chemistry and biology of reactive oxygen species. Nat. Chem. Biol. *4*, 278–286.

Woolley, J.F., Stanicka, J., and Cotter, T.G. (2013). Recent advances in reactive oxygen species measurement in biological systems. Trends Biochem. Sci. *38*, 556–565.

Xu, S., and Chisholm, A.D. (2014). C. elegans epidermal wounding induces a mitochondrial ROS burst that promotes wound repair. Dev. Cell *31*, 48–60.

Yee, C., Yang, W., and Hekimi, S. (2014). The intrinsic apoptosis pathway mediates the pro-longevity response to mitochondrial ROS in C. elegans. Cell *157*, 897–909.

Yun, J., and Finkel, T. (2014). Mitohormesis. Cell Metab. *19*, 757–766.

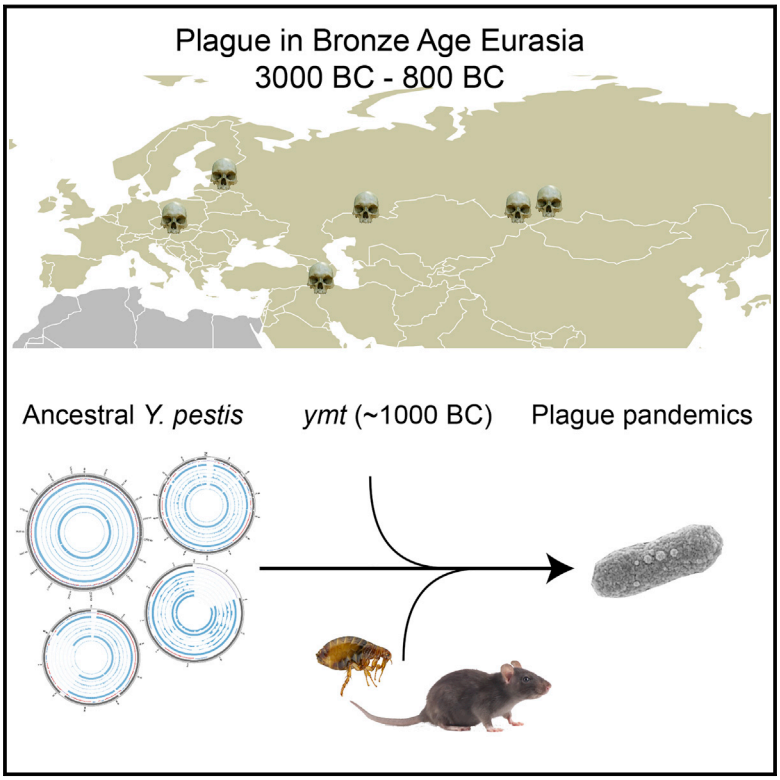
Zarse, K., Schmeisser, S., Groth, M., Priebe, S., Beuster, G., Kuhlow, D., Guthke, R., Platzer, M., Kahn, C.R., and Ristow, M. (2012). Impaired insulin/IGF1 signaling extends life span by promoting mitochondrial L-proline catabolism to induce a transient ROS signal. Cell Metab. *15*, 451–465.

Article

Cell

Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago

Graphical Abstract



Authors

Simon Rasmussen, Morten Erik Allentoft, Kasper Nielsen, ..., Rasmus Nielsen, Kristian Kristiansen, Eske Willerslev

Correspondence

ewillerslev@snm.ku.dk

In Brief

The plague-causing bacteria *Yersinia pestis* infected humans in Bronze Age Eurasia, three millennia earlier than any historical records of plague, but only acquired the genetic changes making it a highly virulent, flea-borne bubonic strain ~3,000 years ago.

Highlights

- *Yersinia pestis* was common across Eurasia in the Bronze Age
- The most recent common ancestor of all *Y. pestis* was 5,783 years ago
- The *ymt* gene was acquired before 951 cal BC, giving rise to transmission via fleas
- Bronze Age *Y. pestis* was not capable of causing bubonic plague

Rasmussen et al., 2015, Cell *163*, 571–582
October 22, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.cell.2015.10.009

CellPress

Article

Cell

Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago

Simon Rasmussen,[1,18] Morten Erik Allentoft,[2,18] Kasper Nielsen,[1] Ludovic Orlando,[2] Martin Sikora,[2] Karl-Göran Sjögren,[3] Anders Gorm Pedersen,[1] Mikkel Schubert,[2] Alex Van Dam,[1] Christian Moliin Outzen Kapel,[4] Henrik Bjørn Nielsen,[1] Søren Brunak,[1,5] Pavel Avetisyan,[6] Andrey Epimakhov,[7] Mikhail Viktorovich Khalyapin,[8] Artak Gnuni,[9] Aivar Kriiska,[10] Irena Lasak,[11] Mait Metspalu,[12] Vyacheslav Moiseyev,[13] Andrei Gromov,[13] Dalia Pokutta,[3] Lehti Saag,[12] Liivi Varul,[10] Levon Yepiskoposyan,[14] Thomas Sicheritz-Pontén,[1] Robert A. Foley,[15] Marta Mirazón Lahr,[15] Rasmus Nielsen,[16] Kristian Kristiansen,[3] and Eske Willerslev[2,17,*]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kongens Lyngby, Denmark

[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark

[3]Department of Historical Studies, University of Gothenburg, 405 30 Gothenburg, Sweden

[4]Section for Organismal Biology, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

[5]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark

[6]Division of Armenology and Social Sciences, Institute of Archaeology and Ethnography, National Academy of Sciences, 0025 Yerevan, Republic of Armenia

[7]Institute of History and Archaeology RAS (South Ural Department), South Ural State University, 454080 Chelyabinsk, Russia

[8]Orenburg Museum of Fine Arts, 460000 Orenburg, Russia

[9]Department of Archaeology and Ethnography, Yerevan State University, 0025 Yerevan, Republic of Armenia

[10]Department of Archaeology, University of Tartu, 51003 Tartu, Estonia

[11]Institute of Archaeology, University of Wrocław, 50-139 Wrocław, Poland

[12]Department of Evolutionary Biology, Estonian Biocentre and University of Tartu, 51010 Tartu, Estonia

[13]Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, 199034 St. Petersburg, Russia

[14]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, 0014 Yerevan, Armenia

[15]Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK

[16]Center for Theoretical Evolutionary Genetics, University of California, Berkeley, California 94720-3140, USA

[17]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

[18]Co-first author

*Correspondence: ewillerslev@snm.ku.dk

http://dx.doi.org/10.1016/j.cell.2015.10.009

SUMMARY

The bacteria *Yersinia pestis* is the etiological agent of plague and has caused human pandemics with millions of deaths in historic times. How and when it originated remains contentious. Here, we report the oldest direct evidence of *Yersinia pestis* identified by ancient DNA in human teeth from Asia and Europe dating from 2,800 to 5,000 years ago. By sequencing the genomes, we find that these ancient plague strains are basal to all known *Yersinia pestis*. We find the origins of the *Yersinia pestis* lineage to be at least two times older than previous estimates. We also identify a temporal sequence of genetic changes that lead to increased virulence and the emergence of the bubonic plague. Our results show that plague infection was endemic in the human populations of Eurasia at least 3,000 years before any historical recordings of pandemics.

INTRODUCTION

Plague is caused by the bacteria *Yersinia pestis* and is being directly transmitted through human-to-human contact (pneumonic plague) or via fleas as a common vector (bubonic or septicemic plague) (Treille and Yersin, 1894). Three historic human plague pandemics have been documented: (1) the First Pandemic, which started with the Plague of Justinian (541–544 AD), but continued intermittently until ~750 AD; (2) the Second Pandemic, which began with the Black Death in Europe (1347–1351 AD) and included successive waves, such as the Great Plague (1665–1666 AD), until the 18th century; (3) the Third Pandemic, which emerged in China in the 1850s and erupted there in a major epidemic in 1894 before spreading across the world as a series of epidemics until the middle of the 20th century (Bos et al., 2011; Cui et al., 2013; Drancourt et al., 1998; Harbeck et al., 2013; Parkhill et al., 2001; Perry and Fetherston, 1997; Wagner et al., 2014). Earlier outbreaks such as the Plague of Athens (430–427 BC) and the Antonine Plague (165–180 AD) may also have occurred, but there is no direct evidence that allows confident attribution to *Y. pestis* (Drancourt and Raoult, 2002; McNeill, 1976).

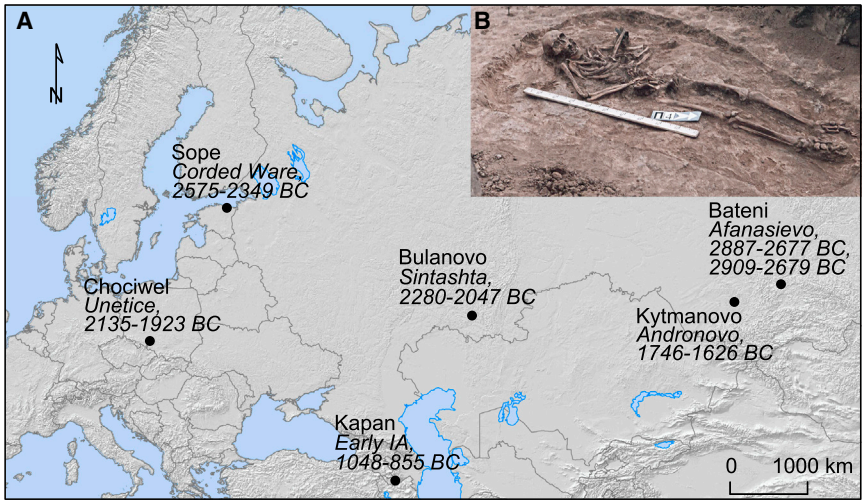


Figure 1. Archaeological Sites of Bronze Age *Yersinia pestis*

(A) Map of Eurasia indicating the position, radiocarbon dated ages and associated cultures of the samples in which *Y. pestis* were identified. Dates are given as 95% confidence interval calendar BC years. IA: Iron Age.

(B) Burial four from Bulanovo site. Picture by Mikhail V. Khalyapin. See also Table S1.

The consequences of the plague pandemics have been well-documented and the demographic impacts were dramatic (Little et al., 2007). The Black Death alone is estimated to have killed 30%–50% of the European population. Economic and political collapses have also been in part attributed to the devastating effects of the plague. The Plague of Justinian is thought to have played a major role in weakening the Byzantine Empire, and the earlier putative plagues have been associated with the decline of Classical Greece and likely undermined the strength of the Roman army.

Molecular clock estimates have suggested that *Y. pestis* diversified from the more prevalent and environmental stress-tolerant, but less pathogenic, enteric bacterium *Y. pseudotuberculosis* between 2,600 and 28,000 years ago (Achtman et al., 1999, 2004; Cui et al., 2013; Wagner et al., 2014). However, humans may potentially have been exposed to *Y. pestis* for much longer than the historical record suggests, though direct molecular evidence for *Y. pestis* has not been obtained from skeletal material older than 1,500 years (Bos et al., 2011; Wagner et al., 2014). The most basal strains of *Y. pestis* (0.PE7 clade) recorded to date were isolated from the Qinghai-Tibet Plateau in China in 1961–1962 (Cui et al., 2013).

We investigated the origin of *Y. pestis* by sequencing ancient bacterial genomes from the teeth of Bronze Age humans across Europe and Asia. Our findings suggest that the virulent, flea-borne *Y. pestis* strain that caused the historic bubonic plague pandemics evolved from a less pathogenic *Y. pestis* lineage infecting human populations long before recorded evidence of plague outbreaks.

RESULTS

Identification of *Yersinia pestis* in Bronze Age Eurasian Individuals

We screened c. 89 billion raw DNA sequence reads obtained from teeth of 101 Bronze Age individuals from Europe and Asia (Allentoft et al., 2015) and found that seven individuals carried sequences resembling *Y. pestis* (Figure 1, Table S1, Supplemental Experimental Procedures). Further sequencing allowed us to assemble the *Y. pestis* genomes to an average depth of 0.14–29.5X, with 12%–95% of the positions in the genome covered at least once (Table 1, Table S2, S3, and S4). We also recovered the sequences of the three plasmids pCD1, pMT1, and pPCP1 (0.12 to 50.3X in average depth) the latter two of which are crucial for distinguishing *Y. pestis* from its highly similar ancestor *Y. pseudotuberculosis* (Table 1, Figure 2, Table S3) (Bercovier et al., 1980; Chain et al., 2004; Parkhill et al., 2001). The host individuals from which *Y. pestis* was recovered belong to Eurasian Late Neolithic and Bronze Age cultures (Allentoft et al., 2015), represented by the Afanasievo culture in Altai, Siberia (2782 cal BC, 2794 cal BC, n = 2), the Corded Ware culture in Estonia (2462 cal BC, n = 1), the Sintashta culture in Russia (2163 cal BC, n = 1), the Unetice culture in Poland (2029 cal BC, n = 1), the Andronovo culture in Altai, Siberia (1686 cal BC, n = 1), and an early Iron Age individual from Armenia (951 cal BC, n = 1) (Table S1).

Authentication of *Yersinia pestis* Ancient DNA

Besides applying standard precautions for working with ancient DNA (Willerslev and Cooper, 2005), the authenticity of our findings are supported by the following observations: (1) The *Y. pestis* sequences were identified in significant amounts in shotgun data from eight of 101 samples, showing that this finding is not due to a ubiquitous contaminant in our lab or in the reagents. Indeed, further analysis showed that one of these eight was most likely not *Y. pestis*. We also sequenced all negative DNA extraction controls and found no signs of *Y. pestis* DNA in these (Table S3). (2) Consistent with an ancient origin, the *Y. pestis* reads were highly fragmented, with average read lengths of 43–65 bp (Table S3) and also displayed clear signs of C-T deamination damage at the 5′ termini typical of ancient DNA (Figure 3, Figure S1). Because the plasmids are central for discriminating between *Y. pestis* and *Y. pseudotuberculosis*, we tested separately for DNA damage patterns for the chromosome and for each of the plasmids. For the seven samples, we observe similar patterns of DNA damage for chromosome and plasmid sequences (Figure 3, Figure S1). (3) We observe correlated DNA degradation patterns when comparing DNA degradation in the *Y. pestis* sequences and the human sequences from the host individual. Given that DNA decay can be described as a rate process (Allentoft et al., 2012), this suggests that the DNA molecules of the pathogen and the human host have a similar age (Figure 3, Figure S1, Table S3 and Supplemental

Table 1. Overview of the *Y. pestis* Containing Samples

Sample	Country	Site	Culture	Date (cal BC)	CO92	pMT1	pPCP1	pCD1
RISE00	Estonia	Sope	Corded Ware	2575–2349	0.39	0.36	1.40	0.66
RISE139	Poland	Chociwel	Unetice	2135–1923	0.14	0.24	0.76	0.28
RISE386	Russia	Bulanovo	Sintashta	2280–2047	0.82	0.96	1.12	1.60
RISE397	Armenia	Kapan	EIA	1048–885	0.25	0.40	6.88	0.50
RISE505	Russia	Kytmanovo	Andronovo	1746–1626	8.73	9.15	34.09	17.46
RISE509	Russia	Afanasievo Gora	Afanasievo	2887–2677	29.45	16.96	31.22	50.32
RISE511	Russia	Afanasievo Gora	Afanasievo	2909–2679	0.20	0.24	1.19	0.60

The dating is direct AMS dating of bones and teeth and is given as 95% confidence interval calendar BC years (details are given in Table S1). The columns CO92, pMT1, pPCP1 and pCD1 correspond to sequencing depth. Additional information on the archaeological sites and mapping statistics can be found in the Supplemental Experimental Procedures and Table S1, S2, and S3. EIA: Early Iron Age, AMS: Accelerator Mass Spectrometry.

Experimental Procedures). (4) Because of the high sequence similarity between *Y. pestis* and *Y. pseudotuberculosis*, we mapped all reads both to the *Y. pestis* CO92 and to the *Y. pseudotuberculosis* IP32953 reference genomes (Chain et al., 2004). Consistent with being *Y. pestis*, the seven investigated samples displayed more reads matching perfectly (edit distance = 0) toward *Y. pestis* (Figure 3, Figure S2). One sample (RISE392) was most likely not *Y. pestis* based on this criterion. (5) A naive Bayesian classifier trained on known genomes predicts the seven samples to be *Y. pestis* with 100% posterior probability, while RISE392 is predicted to have 0% probability of being *Y. pestis* (Figure S2, Table S3). (6) If the DNA was from other organisms than *Y. pestis*, we would expect the reads to be more frequently associated with either highly conserved or low-complexity regions. However, we find the reads to be distributed across the entire genome (Figure S2), and comparison of actual coverage versus the coverage that would be expected from read length distributions and mappability of the reference sequences are also in agreement for the seven samples (Figure 3). (7) In a maximum likelihood phylogeny, the recovered *Y. pestis* genomic sequences of RISE505 and RISE509 are clearly within the *Y. pestis* clade and basal to all contemporary *Y. pestis* strains (Figure 4) (see below).

The Phylogenetic Position of the Bronze Age *Yersinia pestis* Strains

To determine the phylogenetic positions of the two high coverage ancient *Y. pestis* strains, RISE505 (Andronovo culture 1686 cal BC, 8.7X) and RISE509 (Afanasievo culture, 2746 cal BC, 29.7X), we mapped the reads, together with reads from strains of *Yersinia similis* (n = 5), *Y. pseudotuberculosis* (n = 25), and *Y. pestis* (n = 139), to the *Y. pseudotuberculosis* reference genome (IP32953). Only high confidence positions were extracted. To assess whether the individuals were infected with multiple strains of *Y. pestis* we investigated the genotype heterozygosity levels of the ancient genomes and found no indications of mixed infection (Figure S3). There was no decay in Linkage Disequilibrium (LD) across the chromosome (Figure S3), indicating no detectable recombination among strains. We therefore used RAxML (Stamatakis, 2014) to construct a Maximum Likelihood phylogeny from a supermatrix concatenated from 3,141 genes and a total of 3.14 Mbp (Figure 4). This contrasts with earlier phylogenies (Bos et al., 2011; Cui et al., 2013; Morelli et al., 2010; Wagner et al., 2014), which were based on less than 2,300 nucleotides that were ascertained to be variable in *Y. pestis*, likely leading to lower statistical accuracy than with whole-genome analyses. Furthermore, the use of SNPs ascertained to be variable in *Y. pestis* would downwardly bias estimates of branch lengths in *Y. pseudotuberculosis* and lead to underestimates of the *Y. pestis* versus *Y pseudotuberculosis* divergence time, as seen in the branch length of the *Y. pestis* clade to *Y. pseudotuberculosis* (Figure S3). The topology of our whole genome tree shows *Y. pestis* as a monophyletic group within *Y. pseudotuberculosis* with RISE505 and RISE509 (Figure 4A, black arrow, Figure S4) clustered together within the *Y. pestis* clade. The *Y. pestis* sub-tree topology (Figure 4B, Figure S4) is similar to that reported previously (Bos et al., 2011; Cui et al., 2013; Morelli et al., 2010; Wagner et al., 2014), but with the two ancient strains (RISE505 and RISE509) falling basal to all other known strains of *Y. pestis* (100% bootstrap support).

Determination of *Yersinia pestis* Divergence Dates

To determine the dates for the most recent common ancestor (MRCA) of *Y. pestis* and *Y. pseudotuberculosis,* and for all known *Y. pestis* strains, we used a Bayesian Markov Chain Monte Carlo approach implemented in BEAST2 (Bouckaert et al., 2014) on a subset of the supermatrix. We estimated the MRCA of *Y. pestis* and *Y. pseudotuberculosis* to be 54,735 years ago (95% HPD [highest posterior density] interval: 34,659–78,803 years ago) (Figure 4C, Figure S5, Table S5), which is about twice as old compared to previous estimates of 2,600–28,000 years ago (Achtman et al., 1999, 2004; Cui et al., 2013; Wagner et al., 2014). Additionally, we estimated the age of the MRCA of all known *Y. pestis* to 5,783 years ago (95% HPD interval: 5,021–7,022 years ago). This is also significantly older and with a much narrower confidence interval than previous findings of 3,337 years ago (1,505–6,409 years ago) (Cui et al., 2013).

Bronze Age *Yersinia pestis* Strains Lacking Yersinia Murine Toxin

For the high-depth ancient *Y. pestis* genomes, we investigated the presence of 55 genes that have been associated with the virulence of *Y. pestis* (Figure 5A, Table S6). We found all virulence genes to be present, except the Yersinia murine toxin (*ymt*) gene that is located at 74.4–76.2 kb on the pMT1 plasmid (Figure 2C, arrow 1). The *ymt* gene encodes a phospholipase D that protects

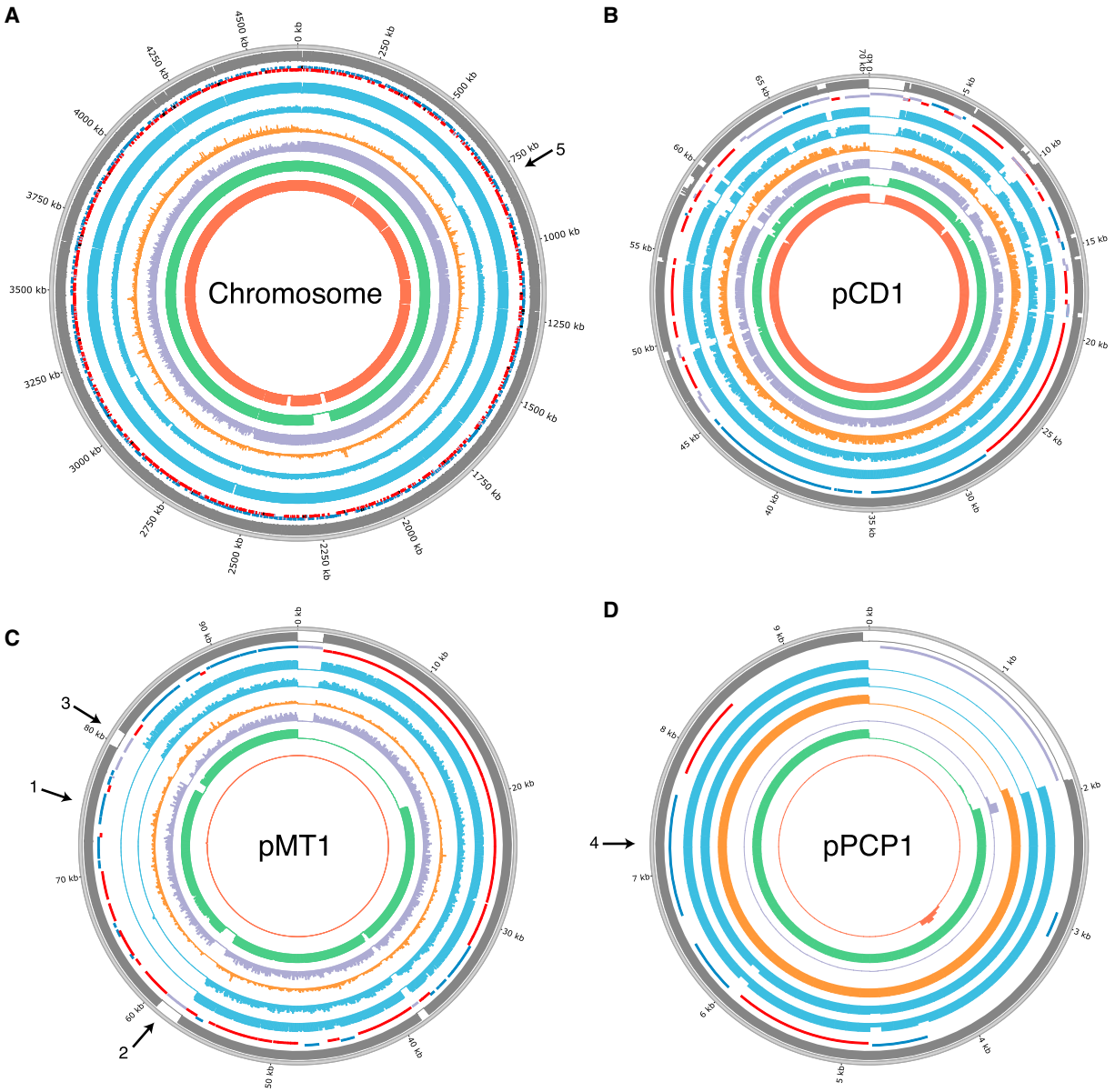


Figure 2. *Y. pestis* Depth of Coverage Plots

(A–D) Depth of coverage plots for (A) CO92 chromosome, (B) pCD1, (C) pMT1, (D) pPCP1. Outer ring: Mappability (gray), genes (RNA: black, transposon: purple, positive strand: blue, negative strand: red), RISE505 (blue), RISE509 (blue), Justinian plague (orange), Black Death plague (purple), modern *Y. pestis* D1982001 (green), *Y. pseudotuberculosis* IP32881 (red) sample. The modern *Y. pestis* and *Y. pseudotuberculosis* samples are included for reference. The histograms show sequence depth in 1 kb windows for the chromosome and 100 bp windows for the plasmids with a max of 20X depth for each ring. Arrow 1: *ymt* gene, arrow 2: transposon at start of missing region on pMT1, arrow 3: transposon at end of missing region on pMT1, arrow 4: *pla* gene, arrow 5: missing flagellin region on chromosome. The plots were generated using Circos (Krzywinski et al., 2009). See also Tables S2, S3 and S8.

Y. pestis inside the flea gut, thus enabling this enteric bacteria to use an arthropod as vector; it further allows for higher titers of *Y. pestis* and higher transmission rates (Hinnebusch, 2005; Hinnebusch et al., 2002). When investigating all seven samples for the presence of *ymt*, we identified a 19 kb region (59–78 kb, Figure 2C arrow 2–3, Figure 5B) to be missing except in the youngest sample (RISE397, 951 cal BC) (Figure 5B, Table S7). We find this region to be present in all other published *Y. pestis* strains (modern and ancient), except three strains (5761, 945, and CA88) that are lacking the pMT1 plasmid completely.

Although larger sample sizes are needed for confirmation, our data indicate that the *ymt* gene was not present in *Y. pestis* before 1686 cal BC (n = 6), while after 951 cal BC, it is found in 97.8% of the strains (n = 140), suggesting a late and very rapid spread of *ymt*. This contrasts with previous studies arguing that the *ymt* gene was acquired early in *Y. pestis* evolution due

to its importance in its life cycle (Carniel, 2003; Hinnebusch, 2005; Hinnebusch et al., 2002; Sun et al., 2014). Interestingly, we identified two transposase elements flanking the missing 19 kb region, confirming that the *ymt* gene was acquired through horizontal gene transfer, as previously suggested (Lindler et al., 1998). Moreover, it has recently been shown that the transmission of *Y. pestis* by fleas is also dependent on loss of function mutations in the *pde2*, *pde3*, and *rcsA* genes (Sun et al., 2014). The RISE509 sample carries the promoter mutation of *pde3* and the functional *pde2* and *rcsA* alleles (Figure S6). In combination with the absence of *ymt*, these results strongly suggest that the ancestral *Y. pestis* bacteria in these early Bronze Age individuals were not transmitted by fleas.

Native Plasminogen Activator Gene Present in Bronze Age *Yersinia pestis*

Another hallmark gene of *Y. pestis* pathogenicity is the plasminogen activator gene *pla* (omptin protein family), located on the pPCP1 plasmid (6.6–7.6 kb). The gene facilitates deep tissue invasion and is essential for development of both bubonic and pneumonic plague (Sebbane et al., 2006; Sodeinde et al., 1992; Zimbler et al., 2015). We identify the gene in six of the seven genomes, but not in RISE139, the sample with the lowest overall depth of coverage (0.75X on pPCP1) (Figure 2D, arrow 4, Table S6). Recently, it has been proposed that pPCP1 was acquired after the branching of the 0.PE2 clade (Zimbler et al., 2015); however, we identified pPCP1 in our samples, including in the 0.PE7 clade (strains 620024 and CMCC05009), which diverged prior to the common ancestor of the 0.PE2 lineage (Figure 4B, Figure 5A). This shows that pPCP1 and *pla* likely were present in the most basal *Y. pestis* (RISE509), suggesting that the 0.PE2 strains lost the pPCP1 plasmid. Interestingly, three 2.ANT3 strains (5761, CMCC64001, and 735) are also missing the *pla* gene, indicating that the loss of pPCP1 occurred more than once in the evolutionary history of *Y. pestis*.

Additionally, we investigated whether RISE397, RISE505, and RISE509 had the isoleucine to threonine mutation at amino acid 259 in the Pla protein. This mutation has been shown to be essential for developing bubonic, but not pneumonic, plague (Zimbler et al., 2015). We found that these samples, in agreement with their basal phylogenetic position, carry the ancestral isoleucine residue. However, we also identified a valine to isoleucine mutation at residue 31 for RISE505 (1686 cal BC) and RISE509 (2746 cal BC). This mutation was not found in any of the other 140 *Y. pestis* strains, but was present in other omptin proteins, such as *Escherichia coli* and *Citrobacter koseri*, and very likely represents the ancestral *Y. pestis* state. The youngest of the samples, RISE397 (951 cal BC) carries the derived isoleucine residue, showing that this mutation, similar to the acquisition of *ymt*, was only observed after 1686 cal BC.

An alternative explanation to the acquisition of *ymt* and the *pla* I259T mutation, given the disparate geographical locations of our samples, could be that the Armenian strain (RISE397, 951 cal BC) containing *ymt* and the isoleucine residue in *pla* had a longer history in the Middle East and experienced an expansion during the 1st millennium BC. This would have led to its export to Eurasia and presumably the extinction of the other more ancestral and less virulent *Y. pestis* strains.

Different Region 4 Present in the Ancestral *Yersinia pestis*

Besides the 55 pathogenicity genes, we also investigated the presence of different region 4 (DFR4) that contains several genes with potential role in *Y. pestis* virulence (Radnedge et al., 2002). This region was reported as present in the Plague of Justinian and Black Death strains, having been lost in the CO92 reference genome (from the Third Pandemic) (Chain et al., 2004; Wagner et al., 2014). Consistent with the ancestral position of our samples, we find evidence that the region is present in all of our seven samples (Figure S6).

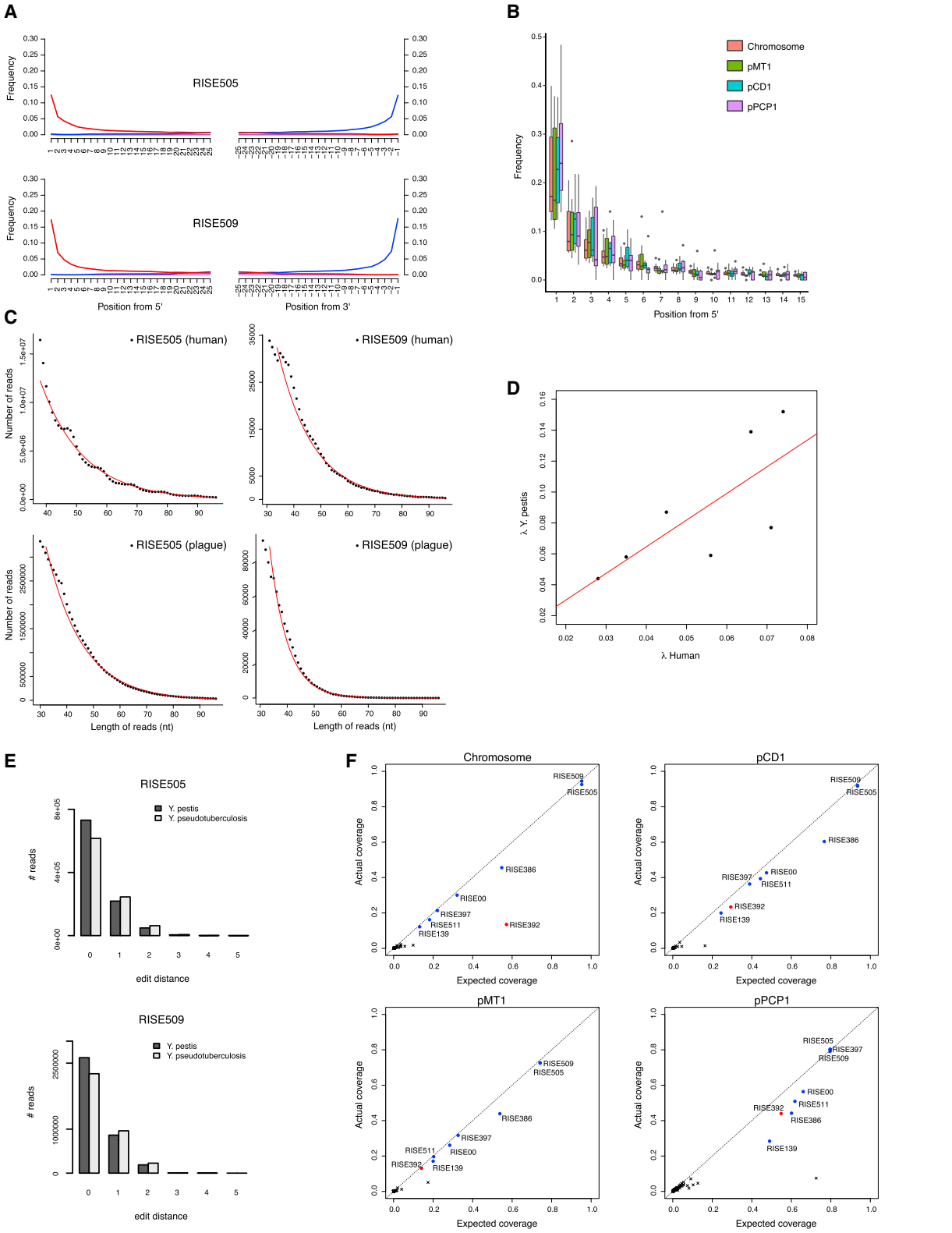
Yersinia pestis flagellar Frameshift Mutation Absent in Bronze Age Strains

Another important feature of *Y. pestis* is the ability to evade the mammalian immune system. Flagellin is a potent initiator of the mammalian innate immune system (Hayashi et al., 2001). *Y. pseudotuberculosis* is known to downregulate expression of flagellar systems in a temperature-dependent manner, and none of the known *Y. pestis* strains express flagellin due to a frameshift mutation in the *flhD* regulatory gene (Minnich and Rohde, 2007). However, we do not find this mutation in either RISE505 or RISE509, suggesting that they have fully functional *flhD* genes and that the loss of function occurred after 2746 cal BC. Interestingly, the youngest of these two *Y. pestis* genomes (RISE505, 1686 cal BC) shows partial loss of one of the two flagella systems (758–806 kb), with 39 of 49 genes deleted (Figure 2A, arrow 5, Table S8). This deletion was not found in any of the other *Y. pestis* samples (n = 147). This may point to selective pressure on ancestral *Y. pestis* when emerging as a mammalian pathogen, yielding variably adaptive strains.

DISCUSSION

Our calibrated molecular clock pushes the divergence dates for the early branching of *Y. pestis* back to 5,783 years ago, an additional 2,000 years compared to previous findings (Table S5, Figure S5) (Cui et al., 2013; Morelli et al., 2010). Furthermore, using the temporally stamped ancient DNA data, we are able to derive a time series for the molecular acquisition of the pathogenicity elements and immune avoidance systems that facilitated the evolution from a less virulent bacteria with zoonotic potential, such as *Y. pseudotuberculosis*, to one of the most deadly bacteria ever encountered by humans (Figure 6).

From our findings, we conclude that the ancestor of extant *Y. pestis* strains was present by the end of the 4th millennium BC and was widely spread across Eurasia from at least the early 3rd millennium BC. The occurrence of plague in the Bronze Age Eurasian individuals we sampled (7 of 101) indicates that plague infections were common at least 3,000 years earlier than recorded historically. However, based on the absence of crucial virulence genes, unlike the later *Y. pestis* strains that were responsible for the first to third pandemics, these ancient ancestral *Y. pestis* strains likely did not have the ability to cause bubonic plague, only pneumonic and septicemic plague. These early plagues may have been responsible for the suggested population declines in the late 4th millennium BC and the early 3rd millennium BC (Hinz et al., 2012; Shennan et al., 2013).



(legend on next page)

It has recently been demonstrated by ancient genomics that the Bronze Age in Europe and Asia was characterized by large-scale population movements, admixture, and replacements (Allentoft et al., 2015; Haak et al., 2015), which accompanied profound and archaeologically well-described social and economic changes (Anthony, 2007; Kristiansen and Larsson, 2005). In light of our findings, it is plausible that plague outbreaks could have facilitated—or have been facilitated by—these highly dynamic demographic events. However, our data suggest that *Y. pestis* did not fully adapt as a flea-borne mammalian pathogen until the beginning of the 1st millennium BC, which precipitated the historically recorded plagues.

EXPERIMENTAL PROCEDURES

Samples and Archaeological Sites

We initially re-analyzed the data from Allentoft et al. (Allentoft et al., 2015) and identified *Y. pestis* DNA sequences in 7 of the 101 individuals. Descriptions of the archaeological sites are given in Supplemental Experimental Procedures and Table S1.

Generation of Additional Sequence Data

In order to increase the depth of coverage on the *Y. pestis* genomes we sequenced more on these seven DNA extracts. Library construction was conducted as in (Allentoft et al., 2015). Briefly, double stranded and blunt-ended DNA libraries were prepared using the NEBNext DNA Sample Prep Master Mix Set 2 (E6070) and Illumina-specific adapters (Meyer and Kircher, 2010). The libraries were "shot-gun" sequenced in two pools on Illumina HiSeq2500 platforms using 100-bp single-read chemistry. We sequenced 32 lanes generating a total of 11.2 billion new DNA sequences for this study. Reads for the seven *Y. pestis* samples are available from ENA: PRJEB10885. Individual sample accessions numbers are available in Table S2.

Creation of Database for Identification of *Y. pestis* Reads

To identify *Y. pestis* reads in the Bronze Age dataset (Allentoft et al., 2015) we first created a database of all previously sequenced *Y. pestis* strains (n = 140), *Y. pseudotuberculosis* strains (n = 30), *Y. similis* strains (n = 5), and a selection of *Y. enterocolitica* strains (n = 4) (Supplemental Experimental Procedures and Table S2). The genomes were either downloaded from NCBI or downloaded as reads and de novo assembled using SPAdes-3.5.0 (Bankevich et al., 2012) with the–careful and–cov-cutoff auto options.

Identification and Assembly of *Y. pestis* From Ancient Samples

Raw reads were trimmed for adaptor sequences using AdapterRemoval-1.5.4 (Lindgreen, 2012). Additionally leading and trailing Ns were removed as well as bases with quality 2 or less. Hereafter, the trimmed reads with a length of at least 30 nt were mapped using bwa mem (local alignment) (Li and Durbin, 2009) to the database of *Y. pestis*, *Y. pseudotuberculosis*, *Y. similis*, and *Y. enterocolitica* mentioned above. Reads with a match to any of the sequences in this database were aligned separately to three different reference genomes: *Yersinia pestis* CO92 genome including the associated plasmids pCD1, pMT1, pPCP1 (Parkhill et al., 2001); *Yersinia pseudotuberculosis* IP32953 including the associated plasmids (Chain et al., 2004); *Yersinia pestis biovar Microtus* 91001 and associated plasmids (Zhou et al., 2004). This alignment was performed using bwa aln (Li and Durbin, 2009) with the seed option disabled for better sensitivity for ancient data, enforcing global alignment of the read to the reference genome. Each sequencing run was merged to library level and duplicates removed using Picard-1.124 (http://broadinstitute.github.io/picard/), followed by merging to per sample alignment files. These files were filtered for a mapping quality of 30 to only retain high quality alignments and the base qualities were re-scaled for DNA damage using MapDamage 2.0 (Jónsson et al., 2013). We defined *Y. pestis* as present in a sample if the mapped depth of the CO92 reference sequences were higher or equal to 0.1X and if the reads covered at least 10% of the chromosome and each of the plasmids. The assembly of Justinian, Black Death, and the modern samples were performed similarly and is described in detail in the Supplemental Experimental Procedures.

Coverage, Depth and Mappability Analyses

We calculated the coverage of the individual sample alignments versus the *Y. pestis* CO92 reference genome using Bedtools (Quinlan and Hall, 2010) and plotted this using Circos (Krzywinski et al., 2009). For the chromosome, the coverage was calculated in 1 kbp windows and for the plasmids in 100 bp windows. Mappability was calculated using GEM-mappability library using a k-mer size of 50, which is similar to the average length of the trimmed and mapped *Y. pestis* reads (average length 43–65 bp). Statistics of the coverage and depth are given in Tables S3 and S4.

DNA Decay Rates

We investigated the molecular degradation signals obtained from the sequencing data. Based on the negative exponential relationship between frequency and sequence length, we estimated for each sample the DNA damage fraction (λ, per bond), the average fragment length (1/ λ), the DNA decay rate (k, per bond per year), and the molecular half-lives of 100 bp fragments (Allentoft et al., 2012). We compared these DNA decay estimates for *Y. pestis* to the decay of endogenous human DNA from the host individuals. If the plague DNA is authentic and ancient, a correlation is expected between the rate of DNA decay in the human host and in *Y. pestis*, because the DNA has been exposed to similar environmental conditions for the same amount of time. See Supplemental Experimental Procedures for additional information.

Figure 3. Authenticity of *Y. pestis* DNA

(A) DNA damage patterns for RISE505 and RISE509. The frequencies of all possible mismatches observed between the *Y. pestis* CO92 chromosome and the reads are reported in gray as a function of distance from 5′ (left panel, first 25 nucleotides sequenced) and distance to 3′ (right panel, last 25 nucleotides). The typical DNA damage mutations C>T (5′) and G>A (3′) are reported in red and blue, respectively.

(B) Ancient DNA damage patterns (n = 7) of the reads aligned to the CO92 chromosome and the *Y. pestis* associated plasmids pMT1, pCD1 and pPCP1. The boxplots show the distribution of C-T damage in the 5′ of the reads. The lower and upper hinges of the boxes correspond to the 25th and 75th percentiles, the whiskers represent the 1.5 inter-quartile range (IQR) extending from the hinges, and the dots represent outliers from these.

(C) DNA fragment length distributions from RISE505 and RISE509 samples representing both the *Y. pestis* DNA and the DNA of the human host. The declining part of the distributions is fitted to an exponential model (red).

(D) Linear correlation (red) between the decay constant in the DNA of the human host and the associated *Y. pestis* DNA extracted from the same individual (R^2 = 0.55, p = 0.055). The decay constant (λ) describes the damage fraction (i.e., the fraction of broken bonds on the DNA strand).

(E) Distribution of edit distance of high quality reads from RISE505 and RISE509 samples mapped to either *Y. pestis* (dark gray) or *Y. pseudotuberculosis* (light gray) reference genomes. The reads have a higher affinity to *Y. pestis* than to *Y. pseudotuberculosis*.

(F) Plots of actual coverage versus expected coverage for the 101 screened samples. Expected coverage was computed taking into account read length distributions, mappable fractions of reference sequences, and the deletions in pMT1 for some of the samples. Samples assumed to contain *Y. pestis* are shown in blue and RISE392 that is classified as not *Y. pestis* appears is shown in red. See also Figure S1 and S2, Table S3.

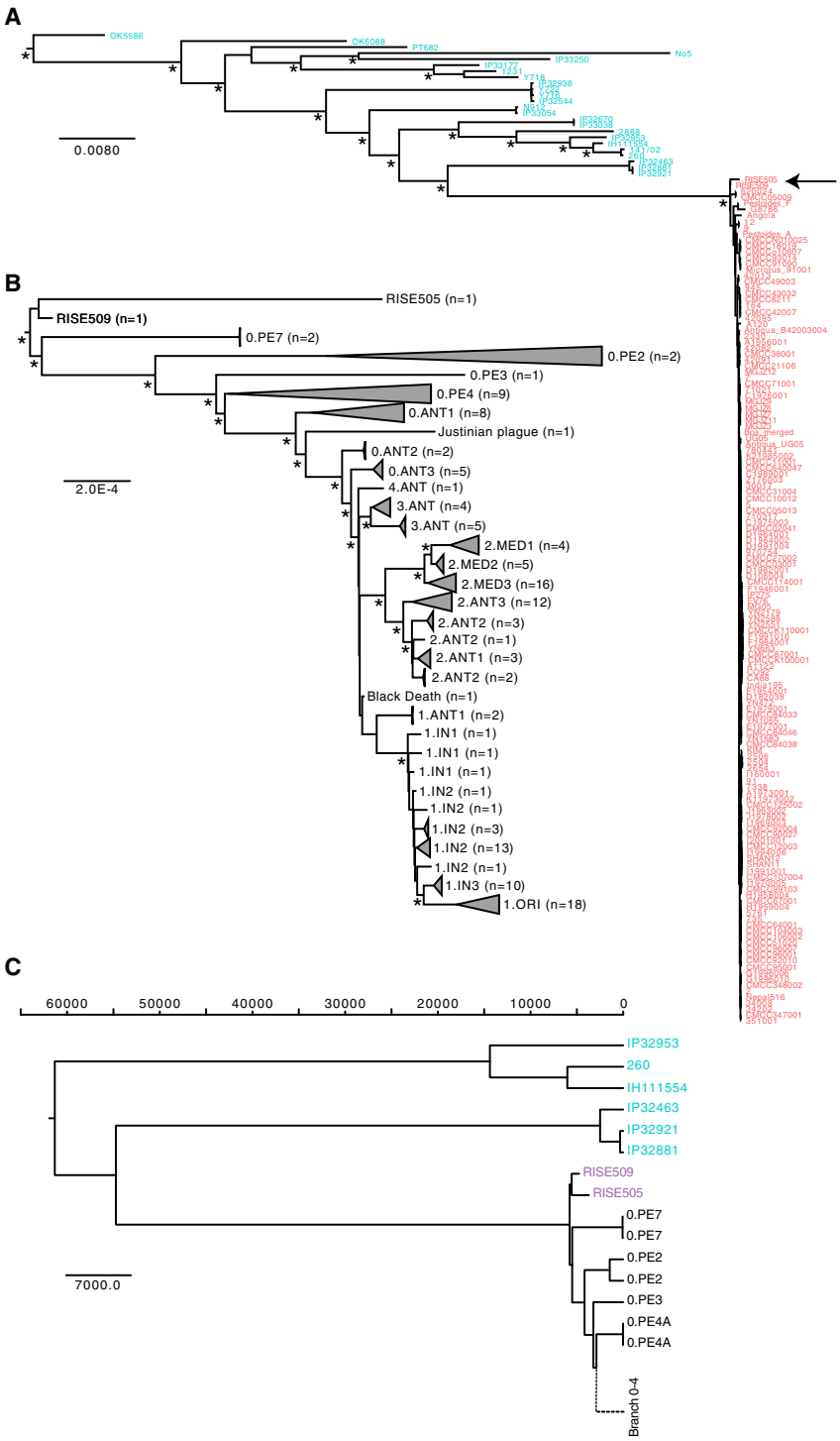


Figure 4. Phylogenetic Reconstructions

(A) Maximum Likelihood reconstruction of the phylogeny of *Y. pseudotuberculosis* (blue) and *Y. pestis* (red). The tree is rooted using *Y. similis* (not shown). The full tree including three additional *Y. pseudotuberculosis* strains (O:15 serovar) can be seen in Figure S4. Major branching nodes within *Y. pseudotuberculosis* with > 95% bootstrap support are indicated with an asterisk and branch lengths are given as substitutions per site.

(B) Maximum Likelihood reconstruction of the phylogeny in (A) showing only the *Y. pestis* clade. The clades are collapsed by population according to branches and serovars, as given in (Achtman et al., 1999, 2004; Cui et al., 2013). See Figure S4 for an uncollapsed tree and Table S2 for details on populations. Nodes with more than 95% bootstrap support are indicated with an asterisk and branch lengths are given as substitutions per site.

(C) BEAST2 maximum clade credibility tree showing median divergence dates. Branch lengths are given as years before the present (see Divergence estimations in Experimental Procedures). Only the *Y. pseudotuberculosis* (blue), the ancient *Y. pestis* samples (magenta) and the most basal branch 0 strains (black) are shown. For a full tree including all *Y. pestis* see Figure S5. See also Figure S3, S4, and S5 and Table S5.

Comparison of Samples to *Y. pestis* and *Y. pseudotuberculosis* Reference Genomes

We used the alignments of several sets of reads (*Y. pestis*, *Y. pseudotuberculosis*, and *Y. similis*) to *Y. pestis* CO92 and the *Y. pseudotuberculosis* IP32953 genomes. Per sample we determined the distribution of edit-distances (mismatches) of the reads versus the particular reference genome. We used these distributions to build a Naive Bayesian classifier to classify whether reads were originating from *Y. pestis*, *Y. pseudotuberculosis*, or *Y. similis*. See Supplemental Experimental Procedures and Table S3.

Expected versus Actual Coverage

We estimated the expected coverage of *Y. pestis* given a specific sequencing depth and correlated that with the actual coverage of a genome per sample. Expected coverage was calculated as

$$c = 1 - \prod_{i=1}^{N} \left(1 - \frac{l_i}{g}\right)^{r_i}$$

where the reads have N different lengths, l_1 to l_N with counts r_1 to r_N. To account for mappability we determined the mappable fraction for each reference sequence using kmers of length 40, 50, and 60, and then used the mappability value with the k-mer length closest to the actual average read length for each sample/reference combination. For more information see Supplemental Experimental Procedures.

Genotyping For Phylogenetic Analyses

Alignments of all strains versus *Y. pseudotuberculosis* IP32953 was used as reference for genotyping the consensus sequences for all samples used in the phylogeny. The samples were genotyped individually using samtools-0.1.18 and bcftools-0.1.17 (Li et al., 2009) and hereafter filtered (Supplemental Experimental Procedures). Based on *Y. pseudotuberculosis* IP32953 gene annotations, the consensus sequences for each gene and sample were extracted. Because of the divergence between *Y. pestis* and *Y. pseudotuberculosis*, a number of gene sequences displayed high rates of missing bases and we removed genes where 20 or more modern *Y. pestis* samples had >10% missingness. This corresponded to a total of 985 genes, leaving data from 3,141 genes that were merged into

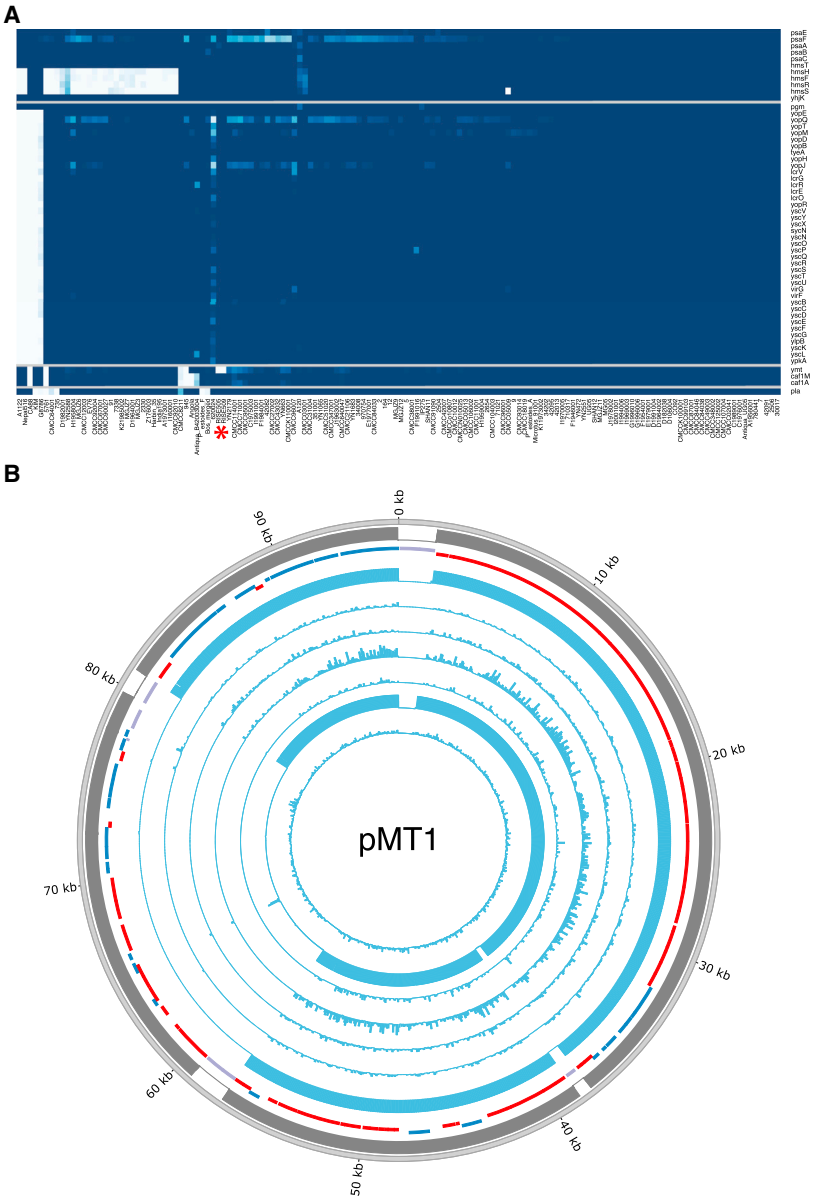


Figure 5. Identification of Virulence Genes

(A) Gene coverage heatmap of 55 virulence genes (rows) in 140 *Y. pestis* strains (columns). Sample ordering is based on hierarchical clustering (not shown) of the gene coverage distributions. RISE505 and RISE509 are marked with a red asterisk. Coloring goes from 0% gene coverage (white) to 100% gene coverage (blue).

(B) Depth of coverage of high quality reads mapping across pMT1. Outer ring is mappability (gray), genes (RNA: black, transposon: purple, positive strand: blue, negative strand: red) and then the RISE samples ordered after direct AMS dating. Sample ordering are RISE509, RISE511, RISE00, RISE386, RISE139, RISE505 and RISE397. See also Figure S6, Tables S2, S6, and S7. AMS: Accelerator Mass Spectrometry.

a supermatrix. We created two different supermatrices, one with *Y. similis*, *Y. pseudotuberculosis*, and *Y. pestis* containing 173 taxa × 3,141 genes that was used for the initial phylogeny (Figure 4A). The second supermatrix consisted of all *Y. pestis* strains and the genomes from the two closest *Y. pseudotuberculosis* clades, which was used for the divergence time estimations.

Phylogenetics

The alignments were partitioned by codon position and analyzed with jmodeltest-2.1.7 (Darriba et al., 2012) to test for the best fitting substitution model. All decision criteria (Akaike, Bayesian, and Decision theory) found the Generalized Time Reversible substitution model with gamma distributed rates, using four rate categories, and a proportion of invariable sites (GTR+G+I) to be the best fit for each of the three codon partitions. To test for recombination across the chromosome we estimated linkage disequilibrium (LD) using 141 *Y. pestis* strains. A total of 482 bi-allelic single nucleotide variations (SNVs), with a minor allele frequency of 5% or higher were extracted. For all pairs of the extracted SNVs, the LD r^2 was calculated using PLINK 1.9 (Chang et al., 2015) and plotted against the physical distance between the pairs. We reconstructed the phylogeny from the codon-partitioned supermatrix using RAxML-8.1.15 (Stamatakis, 2014) with the GTR+G+I substitution model. Bootstraps were performed by generating 100 bootstrap replicates and their corresponding parsimony starting trees using RAxML. Hereafter, a standard Maximum Likelihood inference was run on each bootstrap replicate, and the resulting best trees were merged and drawn on the best ML tree. Initial phylogenies placed the *Y. pestis* Harbin strain with an unusual long branch inside the 1.ORI clade and it was excluded from further analysis. Additionally *Y. pseudotuberculosis* SP93422 (serotype O:15), *Y. pseudotuberculosis* WP-931201 (serotype O:15) and *Y. pseudotuberculosis* Y248 (serotype unknown) was in a clade with long branch lengths and were therefore also omitted (see Figure S4).

Heterozygosity Estimates

We determined heterozygosity by down-sampling the *Y. pestis* bam-files to the same average depth as the corresponding RISE samples, genotyped each of the samples and extracted heterozygote calls with a depth equal to or higher than 10. All transitions were excluded. See Supplemental Experimental Procedures for detailed information.

Divergence Estimations

To date the divergence time for *Y. pestis* and nodes within the *Y. pestis* clade we performed Bayesian Markov Chain Monte Carlo simulations using BEAST-2.3.0 (Bouckaert et al., 2014) and the BEAGLE library v2.1.2 (Ayres et al., 2012). We used the codon-partitioned supermatrix that included the two closest *Y. pseudotuberculosis* clades, with unlinked substitution models, GTR+G+I with eight gamma rate categories and unlinked clock models. Dates were set as years ago with the RISE509, RISE505, Justinian and Black Death samples set to 4,761, 3,701, 1,474, and 667 years ago, respectively. All unknown dates were set to 0 years ago. We followed previous work (Cui et al., 2013; Wagner et al., 2014) and applied a lognormal relaxed clock, assuming a constant population size. We re-rooted the ML tree from RAxML so that the root was placed between the two *Y. pseudotuberculosis* clades (IP32953, 260, IH111554) and (IP32921, IP32881, IP32463) and used this as the starting tree. Based on the ML tree we defined the closets *Y. pseudotuberculosis* clade (IP32921, IP32881, IP32463) and the *Y. pestis* clade as a monophyletic group and defined a uniform prior with 1,000 and 100,000 years as minimum and maximum bounds. We ran 20 independent parallel BEAST chains sampling every 2,000 states for between 52 and 64 million states using a total of 240,000 core hours. The chains were combined using LogCombiner discarding the initial 10 million states as burn-in. The combined post burn-in data represented 961 million states and

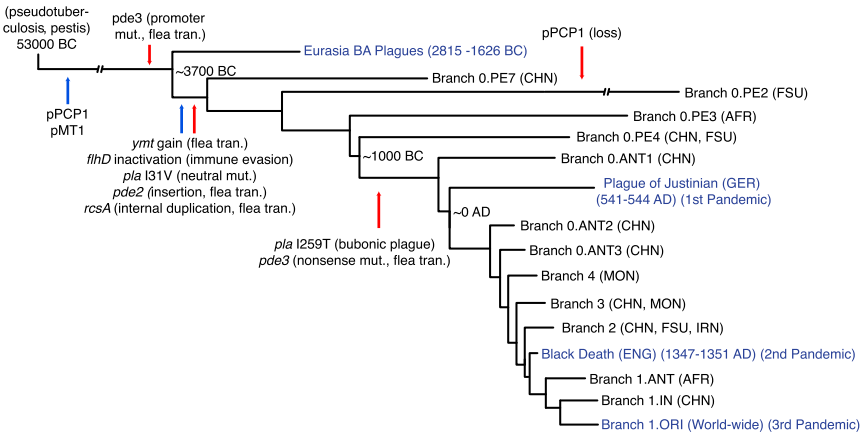


Figure 6. Schematic of *Y. pestis* Evolution

Representation of *Y. pestis* phylogeny and important evolutionary events since divergence from *Y. pseudotuberculosis*. Genetic gains (blue) and genetic loss or loss of function mutations (red) are indicated by arrows. Historical recorded pandemics are indicated in blue text. The calendric years indicates the primary outbreak of the Pandemic. Node dates are median divergence times from the BEAST analysis. The events are based on information from this study and Sun et al., 2014. We used the VCFs generated from all *Y. pestis* samples (n = 142) (Table S2) to verify on which branches the genetic events occurred. The figure is based on current knowledge and is subject to change with addition of new samples. See also Figure S5 and Table S5. BA: Bronze Age, CHN: China, FSU: Former Soviet Union, AFR: Africa, GER: Germany, MON: Mongolia, IRN: Iran, ENG: England, flea tran: flea transmission, mut.: mutation.

the effective sample sizes (ESS) for the posterior was 398, for the TreeHeight 238 and for the MRCA for *Y. pseudotuberculosis* and *Y. pestis* 216. All other parameters had ESS > 125. We then sampled 1/5 of the trees from each chain and combined them for a total of 192,406 trees that were summarized using TreeAnnotator producing a maximum clade credibility tree of median heights. We additionally ran BEAST2 sampling the priors only (and disregarding sequence information) and found the posterior distribution no different than the priors used. It suggests that the posterior distributions recovered when considering full sequence alignments are driven by the sequence information and are not mere by-products of the sampling structure in our dataset (Figure S5).

Analysis of Virulence Associated Genes

To assess the potential virulence of the ancient *Y. pestis* strains, we identified 55 genes previously reported to be associated with virulence of *Y. pestis* (Supplemental Experimental Procedures and Table S6 for details). Based on the alignments to *Y. pestis* CO92 reference genome we determined the fraction of the each gene sequence that was covered by at least one read for each *Y. pestis* sample. Additionally, because the different region 4 (DFR4) (Radnedge et al., 2002) has been associated with virulence, but is not present in the CO92 genome, we used the alignments to *Y. pestis microtus* 91001 to determine the presence of this region (Supplemental Experimental Procedures). We note that the absence of KIM pPCP1 is due to it being missing from the reference genome, but that it has been reported to be present in KIM strains (Hu et al., 1998). The genotypes were generated as described above and the variant call format (VCF) files from these analyses are available at http://www.cbs.dtu.dk/suppl/plague/. For detailed information on genotyping of *pde2*, *pde3*, *rscA*, *pla*, and *flhD* see Supplemental Experimental Procedures.

Identification of the Missing *ymt* Region on pMT1

Most of the regions that were unmapped could be associated with low mappability. However, we identified a region from 59–78 kb on pMT1 that could not be explained by low mappability. From the depth of coverage this region was absent in all of our ancient plague genomes, except for RISE397 (Figure 5). We tested for the significance of this by comparing the distribution of gene depths within and outside of the missing region using the Wilcoxon rank-sum test (Table S7). For all samples except RISE397 the region had a median depth of 0X and the gene depth distributions were significantly different compared to the remaining pMT1 plasmid genes (p values < 1E-9). For the RISE397 sample, the regions had 0.43X and 0.42X median depths and there was no significant difference in the depth of the genes in the two regions (p value 0.77).

ACCESSION NUMBERS

The accession number for the reads for the seven *Y. pestis* samples reported in this paper is ENA: PRJEB10885.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.10.009.

AUTHOR CONTRIBUTIONS

Conceptualization, K-G.S., R.N., K.K. and E.W.; methodology, S.R., M.E.A., A.G.P. and H.B.N.; software, S.R., K.N., M. Sikora, M. Schubert, and A.V.D.; Formal Analysis, S.R., M.E.A., K.N., M. Sikora, A.G.P., A.V.D. and M. Schubert.; Investigation, M.E.A. and K-G.S.; Resources, S.B., P.A., M.V.K., A.E., A. Gnuni, A.K., I.L., M.M., V.M., A. Gromov, D.P., L.S., L.V., L.Y. and T.S-P.; Writing – Original Draft, S.R., M.E.A., K.N., L.O., K-G.S., A.G.P., R.A.F., M.M.L., R.N., K.K. and E.W.; Writing Review & Editing, S.R., M.E.A., K.N., L.O., M. Sikora, K-G.S., A.G.P., A.V.D., C.M.O., R.A.F., M.M.L., R.N., K.K. and E.W.; Visualization, S.R. M.E.A., K-G.S. and A.G.P.; Supervision, L.O., T.S-P., R.N., K.K. and E.W.; Funding Acquisition, K.K. and E.W.

ACKNOWLEDGMENTS

The project was funded by The European Research Council (FP/2007-2013, grant 269442, The Rise), Marie Curie Actions of the European Union (FP7/2007-2013, grant 300554), The Villum Foundation (Young Investigator Programme, grant 10120), University of Copenhagen (KU2016 Programme), The Danish National Research Foundation, and The Lundbeck Foundation. A.V.D. was supported by the National Science Foundation Postdoctoral Research Fellowship in Biology under grant 1306489. S.B. was supported financially by the Novo Nordisk Foundation Grant agreement NNF14CC0001. We thank Jesper Stenderup for technical assistance and want to acknowledge the Danish national supercomputer – Computerome (computerome.cbs.dtu.dk) for the computational resources to perform the BEAST divergence estimations.

Received: August 6, 2015
Revised: September 30, 2015
Accepted: October 2, 2015
Published: October 22, 2015

REFERENCES

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. USA *96*, 14043–14048.

Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A.J., Wagner, D.M., Allender, C.J., Easterday, W.R., et al. (2004). Microevolution

and history of the plague bacillus, Yersinia pestis. Proc. Natl. Acad. Sci. USA *101*, 17837–17842.

Allentoft, M.E., Collins, M., Harker, D., Haile, J., Oskam, C.L., Hale, M.L., Campos, P.F., Samaniego, J.A., Gilbert, M.T.P., Willerslev, E., et al. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc. Biol. Sci. *279*, 4724–4733.

Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.

Anthony, D. (2007). The Horse, The Wheel and Language. How Bronze-Age riders from the Eurasian Steppes Shaped the Modern World (Princeton: Princeton University Press).

Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P., et al. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. *61*, 170–173.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Bercovier, H., Mollaret, H.H., Alonso, J.M., Brault, J., Fanning, G.R., Steigerwalt, A.G., and Brenner, D.J. (1980). Intra- and interspecies relatedness of Yersinia pestis by DNA hybridization and its relationship to Yersinia pseudotuberculosis. Curr. Microbiol. *4*, 225–229.

Bos, K.I., Schuenemann, V.J., Golding, G.B., Burbano, H.A., Waglechner, N., Coombes, B.K., McPhee, J.B., DeWitte, S.N., Meyer, M., Schmedes, S., et al. (2011). A draft genome of Yersinia pestis from victims of the Black Death. Nature *478*, 506–510.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *10*, e1003537.

Carniel, E. (2003). Evolution of pathogenic Yersinia, some lights in the dark. In The Genus Yersinia: Entering the Functional Genomic Era. In The Genus Yersinia, M. Skurnik, J.A. Bengoechea, and K. Granfors, eds. (Boston: Springer US), pp. 3–11.

Chain, P.S.G., Carniel, E., Larimer, F.W., Lamerdin, J., Stoutland, P.O., Regala, W.M., Georgescu, A.M., Vergez, L.M., Land, M.L., Motin, V.L., et al. (2004). Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. USA *101*, 13826–13831.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L.A., Wang, Z., Guo, Z., Xu, L., et al. (2013). Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc. Natl. Acad. Sci. USA *110*, 577–582.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods *9*, 772.

Drancourt, M., and Raoult, D. (2002). Molecular insights into the history of plague. Microbes Infect. *4*, 105–109.

Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O., and Raoult, D. (1998). Detection of 400-year-old Yersinia pestis DNA in human dental pulp: an approach to the diagnosis of ancient septicemia. Proc. Natl. Acad. Sci. USA *95*, 12637–12640.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.

Harbeck, M., Seifert, L., Hänsch, S., Wagner, D.M., Birdsell, D., Parise, K.L., Wiechmann, I., Grupe, G., Thomas, A., Keim, P., et al. (2013). Yersinia pestis DNA from skeletal remains from the 6(th) century AD reveals insights into Justinianic Plague. PLoS Pathog. *9*, e1003349.

Hayashi, F., Smith, K.D., Ozinsky, A., Hawn, T.R., Yi, E.C., Goodlett, D.R., Eng, J.K., Akira, S., Underhill, D.M., and Aderem, A. (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature *410*, 1099–1103.

Hinnebusch, B.J. (2005). The evolution of flea-borne transmission in Yersinia pestis. Curr. Issues Mol. Biol. *7*, 197–212.

Hinnebusch, B.J., Rudolph, A.E., Cherepanov, P., Dixon, J.E., Schwan, T.G., and Forsberg, A. (2002). Role of Yersinia murine toxin in survival of Yersinia pestis in the midgut of the flea vector. Science *296*, 733–735.

Hinz, M., Feeser, I., Sjögren, K.-G., and Müller, J. (2012). Demography and the intensity of cultural activities: an evaluation of Funnel Beaker Societies (4200–2800 cal BC). J. Archaeol. Sci. *39*, 3331–3340.

Hu, P., Elliott, J., McCready, P., Skowronski, E., Garnes, J., Kobayashi, A., Brubaker, R.R., and Garcia, E. (1998). Structural organization of virulence-associated plasmids of Yersinia pestis. J. Bacteriol. *180*, 5192–5202.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

Kristiansen, K., and Larsson, T.B. (2005). The Rise of Bronze Age Society. Travels, Transmissions and Transformations (New York: Cambridge University Press).

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes *5*, 337.

Lindler, L.E., Plano, G.V., Burland, V., Mayhew, G.F., and Blattner, F.R. (1998). Complete DNA sequence and detailed analysis of the Yersinia pestis KIM5 plasmid encoding murine toxin and capsular antigen. Infect. Immun. *66*, 5731–5742.

Little, L.K., Hays, J.N., Morony, M., Kennedy, H.N., Stathakopoulos, D., Sarris, P., Stoclet, A.J., Kulikowski, M., Maddicott, J., Dooley, A., et al. (2007). Plague and the end of antiquity: The pandemic of 541–750 (Cambridge University Press).

McNeill, W.H. (1976). Plagues and Peoples (New York: Anchor Books).

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, pdb.prot5448.

Minnich, S.A., and Rohde, H.N. (2007). A rationale for repression and/or loss of motility by pathogenic Yersinia in the mammalian host. Adv. Exp. Med. Biol. *603*, 298–310.

Morelli, G., Song, Y., Mazzoni, C.J., Eppinger, M., Roumagnac, P., Wagner, D.M., Feldkamp, M., Kusecek, B., Vogler, A.J., Li, Y., et al. (2010). Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat. Genet. *42*, 1140–1143.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L., et al. (2001). Genome sequence of Yersinia pestis, the causative agent of plague. Nature *413*, 523–527.

Perry, R.D., and Fetherston, J.D. (1997). Yersinia pestis–etiologic agent of plague. Clin. Microbiol. Rev. *10*, 35–66.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Radnedge, L., Agron, P.G., Worsham, P.L., and Andersen, G.L. (2002). Genome plasticity in Yersinia pestis. Microbiology *148*, 1687–1698.

Sebbane, F., Jarrett, C.O., Gardner, D., Long, D., and Hinnebusch, B.J. (2006). Role of the Yersinia pestis plasminogen activator in the incidence of distinct

septicemic and bubonic forms of flea-borne plague. Proc. Natl. Acad. Sci. USA *103*, 5526–5530.

Shennan, S., Downey, S.S., Timpson, A., Edinborough, K., Colledge, S., Kerig, T., Manning, K., and Thomas, M.G. (2013). Regional population collapse followed initial agriculture booms in mid-Holocene Europe. Nat. Commun. *4*, 2486.

Sodeinde, O.A., Subrahmanyam, Y.V., Stark, K., Quan, T., Bao, Y., and Goguen, J.D. (1992). A surface protease and the invasive character of plague. Science *258*, 1004–1007.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Sun, Y.-C., Jarrett, C.O., Bosio, C.F., and Hinnebusch, B.J. (2014). Retracing the evolutionary path that led to flea-borne transmission of Yersinia pestis. Cell Host Microbe *15*, 578–586.

Treille, G., and Yersin, A. (1894). La peste bubonique à Hong Kong. VIIIe Congrès Int. D'hygiène Démographie.

Wagner, D.M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J.W., Enk, J., Birdsell, D.N., Kuch, M., Lumibao, C., et al. (2014). Yersinia pestis and the plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect. Dis. *14*, 319–326.

Willerslev, E., and Cooper, A. (2005). Ancient DNA. Proc. Biol. Sci. *272*, 3–16.

Zhou, D., Tong, Z., Song, Y., Han, Y., Pei, D., Pang, X., Zhai, J., Li, M., Cui, B., Qi, Z., et al. (2004). Genetics of metabolic variations between Yersinia pestis biovars and the proposal of a new biovar, microtus. J. Bacteriol. *186*, 5147–5152.

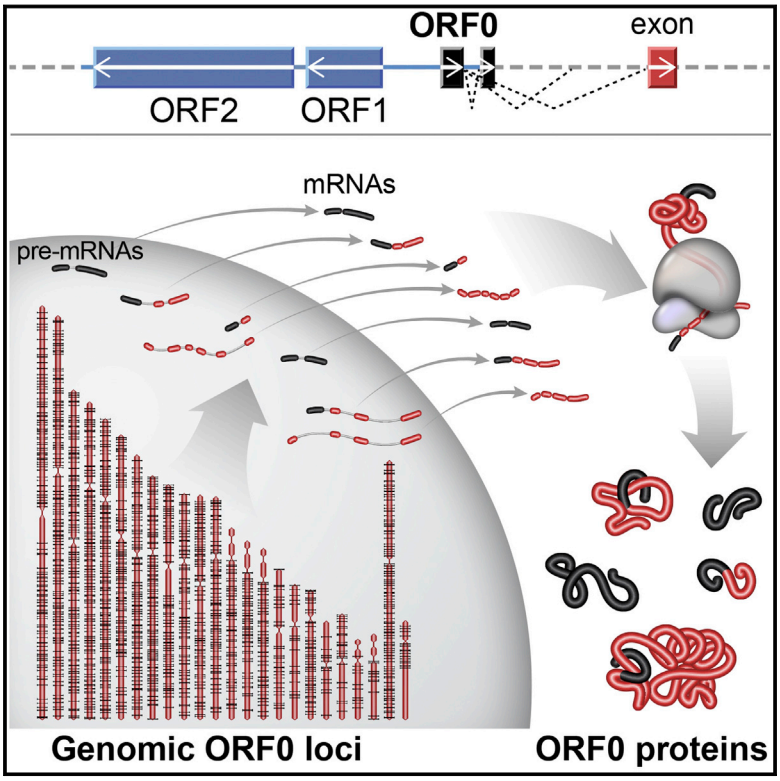
Zimbler, D.L., Schroeder, J.A., Eddy, J.L., and Lathem, W.W. (2015). Early emergence of Yersinia pestis as a severe respiratory pathogen. Nat. Commun. *6*, 7487.

Article

Cell

Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity

Graphical Abstract



Authors

Ahmet M. Denli, Iñigo Narvaiza, Bilal E. Kerman, ..., Tony Hunter, Alan Saghatelian, Fred H. Gage

Correspondence

gage@salk.edu

In Brief

A primate-specific open reading frame, ORF0, is found in the LINE-1 retrotransposons. It not only enhances LINE-1 mobility but also leads to the generation of ORF0-proximal exon fusion proteins, contributing to retrotransposon-mediated diversity.

Highlights

- ORF0 is a primate-specific open reading frame in LINE-1 retrotransposons
- Over 3,000 potential ORF0 loci exist in human and chimpanzee genomes
- ORF0-proximal exon fusion proteins are generated through splicing
- ORF0 influences LINE-1 mobility

Denli et al., 2015, Cell *163*, 583–593
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.025

CellPress

Article

Cell

Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity

Ahmet M. Denli,[1] Iñigo Narvaiza,[1] Bilal E. Kerman,[1] Monique Pena,[1] Christopher Benner,[1] Maria C.N. Marchetto,[1] Jolene K. Diedrich,[5] Aaron Aslanian,[2] Jiao Ma,[3,6] James J. Moresco,[5] Lynne Moore,[1] Tony Hunter,[2,4] Alan Saghatelian,[3] and Fred H. Gage[1,7,8,*]

[1]Laboratory of Genetics
[2]Molecular and Cell Biology Laboratory
[3]Clayton Foundation Laboratories for Peptide Biology
[4]Cancer Center
[5]Mass Spectrometry Center
The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA
[6]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA
[7]Center for Academic Research and Training in Anthropogeny (CARTA)
[8]Kavli Institute for Brain and Mind (KIBM)
University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
*Correspondence: gage@salk.edu
http://dx.doi.org/10.1016/j.cell.2015.09.025

SUMMARY

LINE-1 retrotransposons are fast-evolving mobile genetic entities that play roles in gene regulation, pathological conditions, and evolution. Here, we show that the primate LINE-1 5′UTR contains a primate-specific open reading frame (ORF) in the antisense orientation that we named ORF0. The gene product of this ORF localizes to promyelocytic leukemia-adjacent nuclear bodies. ORF0 is present in more than 3,000 loci across human and chimpanzee genomes and has a promoter and a conserved strong Kozak sequence that supports translation. By virtue of containing two splice donor sites, ORF0 can also form fusion proteins with proximal exons. ORF0 transcripts are readily detected in induced pluripotent stem (iPS) cells from both primate species. Capped and polyadenylated ORF0 mRNAs are present in the cytoplasm, and endogenous ORF0 peptides are identified upon proteomic analysis. Finally, ORF0 enhances LINE-1 mobility. Taken together, these results suggest a role for ORF0 in retrotransposon-mediated diversity.

INTRODUCTION

Transposable elements (TEs) are mobile genetic elements that can alter their chromosomal locations in the host genomes. TEs, first discovered by Barbara McClintock in maize (McClintock, 1950), are abundantly present in nearly all genomes studied to date; they influence gene expression and shape the genomes over evolutionary time (Huang et al., 2012). There are two classes of TEs based on their transposition mechanisms: DNA transposons and retrotransposons. DNA transposons mobilize with a cut-and-paste mechanism, whereas retrotransposons move by copy-and-paste via an RNA intermediate (Kleckner, 1990; Luan et al., 1993). Autonomous elements from both classes are defined as TEs that encode the proteins required for transposition, whereas non-autonomous elements depend on such proteins to be provided in *trans*. In primate genomes, most active TEs belong to the retrotransposon families. Of these, LINE-1 (L1) elements are the only autonomous elements that are currently active (Dewannieux et al., 2003; Hancks et al., 2011) and thus have directly and indirectly contributed to ~30% of the human genome (Lander et al., 2001). At present, the majority of L1 elements are inactive, due to accumulated mutations as well as 5′ truncations that are common during the integration process, thus reducing the number of estimated active elements to ~80 per genome (Brouha et al., 2003). The first active L1 element was isolated through analysis of mutagenic L1 insertions into the factor VIII gene in hemophilia A patients (Dombroski et al., 1991). Since then, retrotransposon germline insertions have been linked to ~100 human diseases (Hancks and Kazazian, 2012).

Intact, active L1s are ~6 kb long and contain a 5′UTR, two open reading frames (ORF1 and ORF2) and a short 3′UTR (Scott et al., 1987). The L1 5′UTR has promoter activity in both the sense and antisense (ASP) directions (Speek, 2001; Swergold, 1990). ORF1 encodes an ~40 kDa RNA-binding protein that is required for L1 transposition (Kolosha and Martin, 1997; Moran et al., 1996). However, ORF1 does not have any significant sequence similarity to known proteins (Goodier et al., 2007). ORF2 is a large protein at ~150 kDa with endonuclease and reverse transcriptase activities (Mathias et al., 1991). These activities, as well as the function of a cysteine-rich region at the C terminus, are important for L1 mobility (Feng et al., 1996; Moran et al., 1996).

Regardless of their ability to mobilize, L1s contribute to transcriptome diversity and gene regulation (Cordaux and Batzer, 2009). Transcription initiated in both directions can extend beyond the L1 sequence but, due to the presence of a polyA signal at the end of the 3′UTR, most sense transcripts end within

the element. However, extensions into the genomic flank are also frequently observed and can lead to 3′ transductions (Moran et al., 1999). Analyses of cloned cDNAs provide evidence of antisense transcripts that are spliced into exons in the neighboring genomic sequences (Macia et al., 2011; Mätlik et al., 2006; Wheelan et al., 2005). Recent studies have focused on specific examples of spliced transcripts with a focus on disease, and a number of L1-driven transcripts have been shown to exist in cancer cells (Cruickshanks and Tufarelli, 2009). In addition to driving genes, antisense transcripts have been linked to chromatin modifications that influence gene expression (Cruickshanks et al., 2013).

A recent analysis of L1s in primates showed that, while ORF1 and ORF2 sequences have been relatively well conserved, acquisition of new 5′UTRs frequently occurred during primate evolution, providing the diversity that resulted in selection of the current 5′UTR (Khan et al., 2006). With the above in mind, we set out to improve our understanding of the properties of the primate L1 5′UTR. Here, we show that the currently active primate L1 5′UTR has well-conserved properties that support translation of an ORF that we have named ORF0. ORF0 is encoded by a primate-specific antisense ORF that lies downstream from the ASP and has a strong, well-conserved Kozak sequence. The gene product of this ORF is predominantly nuclear and localizes to promyelocytic leukemia (PML)-adjacent bodies. ORF0 also has two prominent splice donor (SD) sites at nucleotides 106 and 191 (amino acids 35 and 64) that can act in concert with splice acceptors (SAs) in downstream genomic sequences to generate fusion proteins. ORF0 mRNAs are capped, polyadenylated, associated with ribosomes, and upon immunoaffinity purification, peptides from endogenous ORF0 products can be detected by mass spectrometry. Lastly, overexpression of ORF0 leads to a modest but significant increase in L1 mobility. Thus, we have identified and begun to characterize a third ORF from primate L1 retrotransposons.

RESULTS

Identification of an ORF in the Human Antisense L1 5′UTR

We started by analyzing the antisense 5′UTR for the presence of ORFs that have an upstream promoter, start with ATG, and have a strong Kozak sequence determined by the presence of A/G in position −3 and G at position +4 (Kozak, 1987). Only one potential ORF exists that meets these criteria and, due to its 5′ position with respect to ORF1 and ORF2, we have called it ORF0. ORF0 lies between nucleotides 452–236 from the 5′ end of LINE-1 in the antisense orientation and contains two SD sites (red boxes) within the potential coding sequence (Figure 1A). There are ~781 loci that could encode full-length (FL) ORF0 in the human genome; the consensus sequence for the FL ORF0 protein obtained from these loci is shown in Figure 1A. The chimp ORF0 consensus sequence from ~395 FL ORF0 loci is identical to that of the human.

The previously mapped L1 ASP lies upstream of ORF0, with some overlap (Speek, 2001). This overlap prompted us to check whether the promoter activity resided upstream of the initiator methionine (1st Met) of ORF0. Results from luciferase reporter assays suggested that promoter activity was upstream but not downstream of ORF0 1st Met, and we further mapped a minimal ORF0 promoter of ~150 bp that had similar activity to the previously described L1 ASP (Figure 1B). We also cloned a number of polymorphic ORF0 promoters upstream of luciferase and GFP reporters. While variable, all the tested promoters were active (data not shown). This finding is consistent with previous observations that a high percentage of L1 5′UTRs have antisense promoter activity (Macia et al., 2011). Next, in vitro translation of HA-tagged ORF0 was tested in rabbit reticulocyte lysates and confirmed with western blot analysis (Figure 1C).

To investigate whether this potential ORF could be translated in human cells, we removed the stop codon of ORF0 and cloned it upstream of a promoterless, in-frame GFP coding sequence that lacked the first ATG. Upon transfection, western blot analysis showed that, indeed, the ORF0 promoter and the context around the 1st Met of ORF0 were sufficient to translate the ORF0-GFP fusion protein (Figure 1D).

ORF0 Protein Is Predominantly Nuclear and Present in PML-Adjacent Foci

To analyze the subcellular localization of ORF0, we generated a GFP-tagged ORF0 clone in an L1 context (GFP-ORF0-L1). Since two SD sequences that were often involved in generation of spliced antisense transcripts (Speek, 2001) fell within ORF0, to allow detection of both spliced and unspliced products, GFP was placed at the N terminus but downstream of the Kozak context of ORF0 to minimize any effects on translation initiation (Figure 1E). Western blot analysis confirmed that GFP-ORF0 fusion protein was generated (Figure 1E). Importantly, when the 1st Met of ORF0 was mutated to threonine (M1T), we observed that GFP signal was lost, showing that translation started from the 1st Met of ORF0 (Figure 1F) and ruling out any potential upstream translation initiation. Furthermore, addition of a poly A signal downstream of ORF0 at the end of the L1 did not change protein localization, suggesting that the produced ORF was contained within the L1 and was not a splicing product with the downstream flank (Figure S1A). We also fused ORF0 to mCherry (29% identity to EGFP) and observed a very similar pattern, suggesting that the sequence of the tag was not driving the localization (Figures S1B and S1C). Interestingly, ORF0, but not GFP-alone from the same plasmid backbone, was localized predominantly in nuclear foci in the majority of cells (Figures 1F and S1D–S1F). As predicted by the charge distribution of amino acid residues, the C terminus portion of ORF0 was required for nuclear localization (Figure S1G). Since a number of ORF0 variants may be encoded due to polymorphisms in L1 sequences, we cloned some of these variants and observed that, unless truncated, most localized similarly (data not shown).

Based on the numbers and distribution of foci, we hypothesized that ORF0 localization could be related to PML bodies. PML bodies are nuclear proteinaceous structures often associated with the nuclear matrix and are involved in a wide variety of processes that may influence L1 biology: stress, anti-viral and DNA damage response, transcriptional regulation, heterochromatin, and post-translational protein modifications (Bernardi and Pandolfi, 2007). Indeed, in cells transfected with PML-IV-GFP and mCherry-ORF0, high-magnification imaging

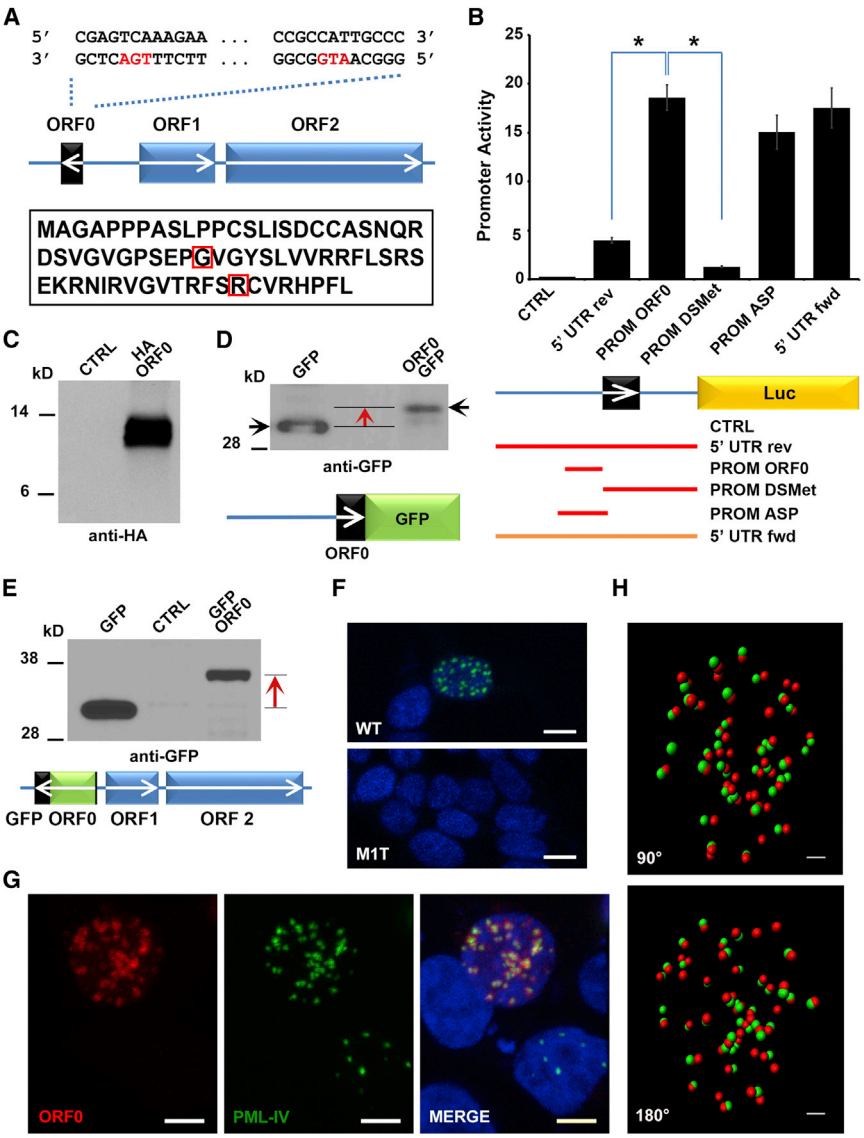


Figure 1. Identification of ORF0 in L1 5′UTR

(A) Location of ORF0 in L1. The start codon ATG and the stop codon TGA are labeled red in the antisense orientation. The positions of splice donor sites within the coding sequence are indicated with red squares. Consensus protein sequence of full-length ORF0 based on ~781 potential ORF0 loci in the human genome.

(B) Upstream ~150 bp region of ORF0 has promoter activity. Luciferase assays were performed to determine promoter activity of the L1 5′UTR regions shown in the panel below the graph. Red and orange lines represent antisense and sense strands, respectively. DSMet refers to downstream of initiator methionine. Data are presented as mean ± SEM. *Denotes $p < 0.05$ significance between indicated groups using t test. CTRL denotes control.

(C) ORF0 can be translated in vitro. HA-tagged ORF0 production was monitored by western blotting.

(D) Production of ORF0-GFP fusion protein was detected by GFP western blot. The C-terminal GFP tagged ORF0 construct driven by the upstream region of ORF0 is shown at the bottom. Black arrows indicate GFP alone and the fusion protein. Red arrow highlights the size shift.

(E) GFP-ORF0 fusion protein was detected by western blot. Design of GFP-ORF0 construct in L1 context. GFP is cloned at the N terminus of ORF0 downstream of the 1st Met and potential Kozak context. Red arrow highlights the size shift in the generated protein.

(F) Translation of GFP-ORF0 is dependent on the ORF0 initiator methionine. Fluorescent detection of ORF0 localization upon transfection of the construct depicted in E into HEK293T cells. WT, wild-type; M1T, initiator methionine to threonine mutant. Scale bar, 10 μm.

(G) Most of ORF0 protein localizes to PML-adjacent nuclear bodies. Confocal imaging of cells transfected with mCherry-ORF0- and GFP-PML-IV-encoding plasmids. Scale bar, 4 μm.

(H) Spot representation of ORF0 (red) and PML (green) foci. Images from 90° and 180° relative to Movie S1 are shown. Scale bar, 1 μm.

See also Figure S1.

showed that ORF0 was present in PML-adjacent foci (Figure 1G). Spot analysis of confocal z series confirmed this observation (Figure 1H; Movie S1).

A Large Number of ORF0 Loci with a Conserved Functional Kozak Context Exist in Primate Genomes

We sought to determine how many loci could potentially encode ORF0 in the human and chimp genomes. Taking splicing into consideration, we scanned these genomes for potential ORF0 loci that are untruncated up to the two commonly used SD sites and have an adjacent GT dinucleotide. Human and chimp genomes have ~3,528 and ~3,299 such loci (of which ~974 and ~745 are species-specific, respectively) that have the potential to splice into the genomic flanks and generate fusion proteins (Figures 2A and 2B). All FL ORF0 loci contain at least one SD and, as a result, they are present in this set. L1 family classification of ORF0 loci are shown in Table S1. Considering insertional polymorphisms within populations and somatic insertions, the number of ORF0 loci may be even larger. Analysis of human and chimp genomes for ORF0 loci revealed a conserved strong Kozak context around the first ATG (Figure 2C). To test the functionality of the consensus wild-type ORF0 Kozak (WT ORF0), we mutated it to an optimal Kozak sequence (OPT) as well as a −3/+4 mutant (MT ORF0). Expression of GFP-ORF0 was comparable between WT ORF0 and OPT, whereas the −3/+4 mutation abolished translational activity (Figures 2D and S2A).

We also extended our ORF0 analysis across mammalian genomes and found ORF0 loci with homology throughout the potential coding sequence, only in the genomes of Catarrhini. Within this parvorder of primates, Old World monkey and ape genomes contain on average ~50 and ~2,500 such ORF0 loci, respectively. Consensus Kozak sequences derived from these

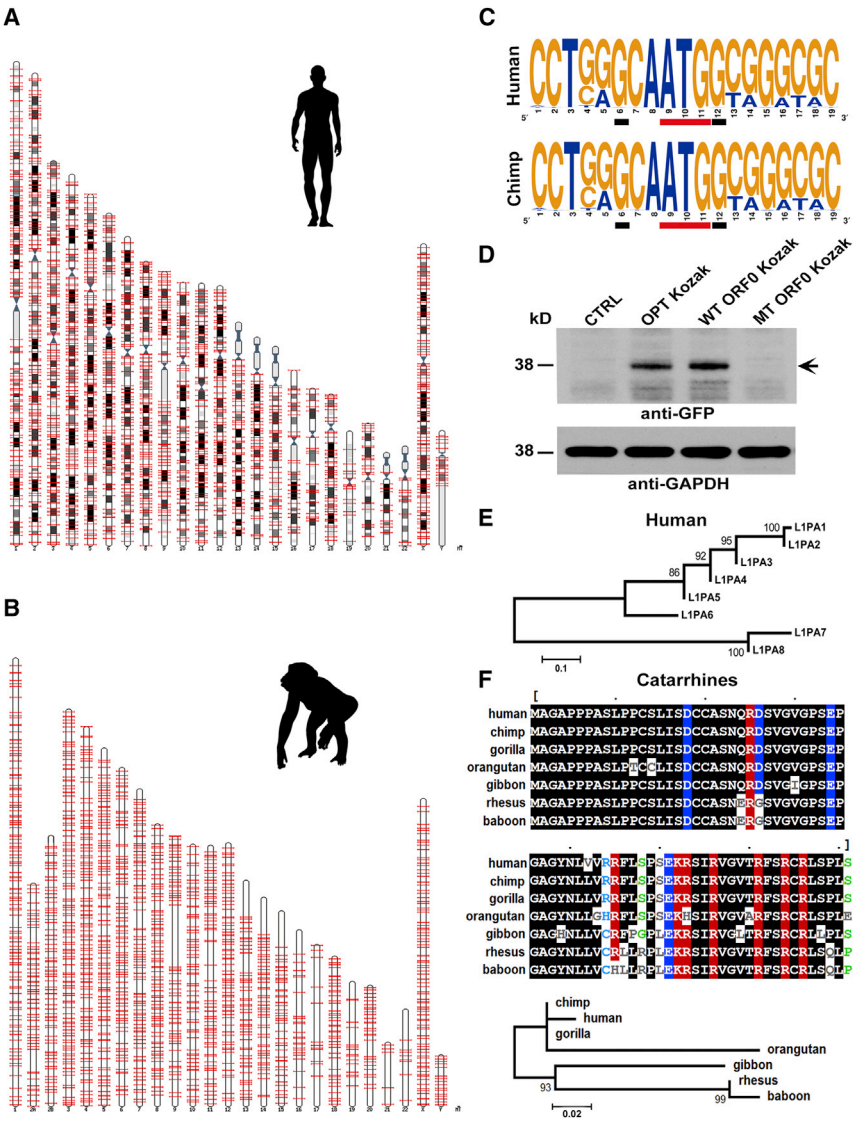


Figure 2. More than 3,000 Potential ORF0 Loci with a Conserved and Functional Kozak Sequence Exist in the Human and Chimp Genomes

(A and B) Chromosomal locations of ORF0 loci in the human and chimp reference genomes. The human and chimp genomes have ~3,528 and ~3,299 loci, respectively, that have the potential to splice into the genomic flanks and generate fusion proteins.

(C) ORF0 loci have a conserved strong Kozak context. Logo of Kozak sequences of ORF0 loci in human and chimp genomes. Start codon is underlined with red, and important nucleotides for translation initiation are underlined with black.

(D) The ORF0 Kozak sequence is functional. Western blot analysis of ORF0-GFP fusions driven by optimal (OPT), wild-type (WT ORF0), and mutant (MT ORF0) Kozak sequences from the GFP-ORF0-L1 construct. Arrow highlights the GFP-ORF0 protein.

(E) Basic phylogenetic analysis of ORF0 sequences in human L1PA families. ORF0 coding sequences were extracted from L1PA family consensus sequences and used in generating the maximum likelihood tree.

(F) Alignment of consensus ORF0 sequences derived from Catarrhini species. Charged residues are labeled in red and blue for positively and negatively charged, respectively. These consensus sequences were used in building the maximum likelihood tree for these primate species.

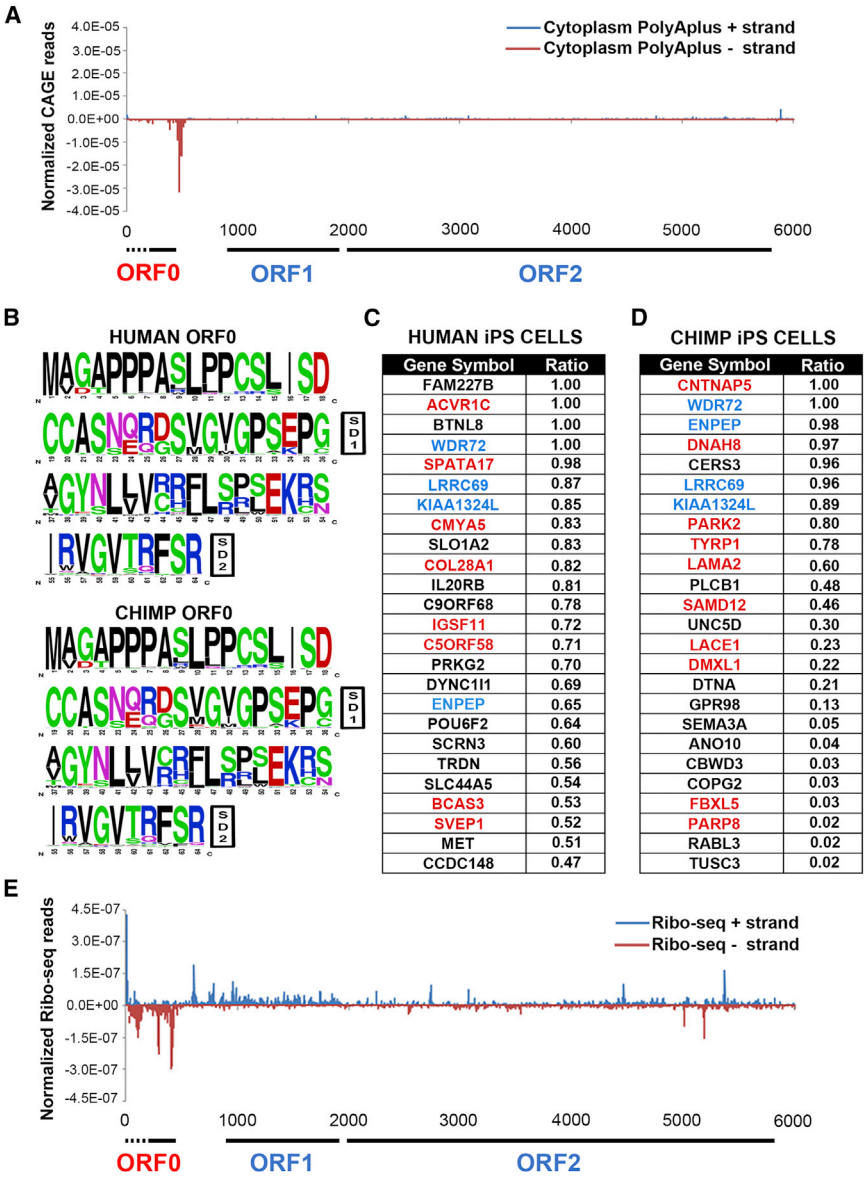
See also Figure S2 and Table S1.

loci suggest that the ORF0 Kozak context is conserved, including the G−3 and G+4 positions (red boxes) (Figure S2B). In New World monkeys, a very small number of ORF0 loci with limited N terminus homology were observed; however, due to the low number, a reliable consensus could not be built and thus these genomes were excluded from further investigation.

We next focused on the ORF0 coding sequences to get a better picture of evolutionary conservation of ORF0 within human L1 families and across primates. The alignments of ORF0 proteins from consensus L1PA1–8 sequences (Khan et al., 2006) are shown in Figure S2C. L1PA1 (that includes L1HS) and L1PA2 have intact SD1 and SD2. L1PA3-L1PA6 families contain a longer ORF0 due to a frameshift after SD2. In L1PA5 and L1PA6, SD1 is mutated but SD2 is conserved (data not shown). L1PA7 and L1PA8 have C termini that are distinct from the other L1PA families and lack SD1 and SD2. The abovementioned variation across L1PA families was recapitulated in the maximum likelihood tree (Figure 2E). Next, we generated consensus ORF0 sequences from the Catarrhines for comparison (Figure 2F). These primates have very similar consensus ORF0 proteins, except for the region between residues ~42 and 50. While all species' consensus ORF0 sequence contains SD2, rhesus and baboons lack SD1 due to a point mutation (Figure S2D). The maximum likelihood tree from the ORF0 sequences of Catarrhine genomes is shown in Figure 2F.

Capped and Polyadenylated ORF0 mRNAs Are Present in the Cytoplasm

One would expect ORF0 to be tightly regulated as a transposable element protein. In addition, short ORFs are technically challenging to uncover (Andrews and Rothnagel, 2014). To determine whether transcription from ORF0 loci could be detected, we turned to transcriptomic data. Cap analysis of gene expression (CAGE) data allow the mapping of transcription start sites (TSSs) and thus make it possible to identify the 5′ end of transcripts that originate from L1 (Faulkner et al., 2009; Shiraki et al., 2003). Our analysis of CAGE data showed that the majority of TSSs for antisense RNAs are upstream of ORF0 1st Met, suggesting that most antisense transcripts could have the capacity



Gene Symbol	Ratio
FAM227B	1.00
ACVR1C	1.00
BTNL8	1.00
WDR72	1.00
SPATA17	0.98
LRRC69	0.87
KIAA1324L	0.85
CMYA5	0.83
SLO1A2	0.83
COL28A1	0.82
IL20RB	0.81
C9ORF68	0.78
IGSF11	0.72
C5ORF58	0.71
PRKG2	0.70
DYNC1I1	0.69
ENPEP	0.65
POU6F2	0.64
SCRN3	0.60
TRDN	0.56
SLC44A5	0.54
BCAS3	0.53
SVEP1	0.52
MET	0.51
CCDC148	0.47

Gene Symbol	Ratio
CNTNAP5	1.00
WDR72	1.00
ENPEP	0.98
DNAH8	0.97
CERS3	0.96
LRRC69	0.96
KIAA1324L	0.89
PARK2	0.80
TYRP1	0.78
LAMA2	0.60
PLCB1	0.48
SAMD12	0.46
UNC5D	0.30
LACE1	0.23
DMXL1	0.22
DTNA	0.21
GPR98	0.13
SEMA3A	0.05
ANO10	0.04
CBWD3	0.03
COPG2	0.03
FBXL5	0.03
PARP8	0.02
RABL3	0.02
TUSC3	0.02

Figure 3. ORF0-Gene Fusion Transcripts Are Expressed in Human and Chimp iPS Cells

(A) Most of the antisense L1 transcription starts upstream of ORF0. Cytoplasmic polyA plus K562 CAGE (ENCODE/RIKEN) reads were mapped to L1HS consensus sequence.

(B) Protein logo of ORF0 loci that are untruncated until splice donor sites in the human and chimp genomes. Sequences from SD1 and SD2 loci are represented as protein sequence logos. Positions of SD1 and SD2 are indicated with black boxes.

(C and D) Table of top 25 protein-coding genes, for which RNA-Seq reads were detected at the splice junction with ORF0 in human and chimp iPS cells. Red-labeled genes have ORF0 fusions due to species-specific L1 insertions. Blue labeling represents genes for which ORF0 fusion transcripts were detected in both human and chimp iPS samples. Transcripts of black-labeled gene fusions are detected only in one species. The ratios of ORF0 isoforms with respect to the total (i.e., ORF0 + annotated gene isoforms) are shown in the ratio column. Table was sorted for ratio from high to low.

(E) Ribosome footprinting data from HEK293T cells were mapped to the L1HS consensus sequence.

See also Figure S3 and Table S2.

to encode ORF0 (Figure S3A). More importantly, ORF0 mRNA could be detected not only in whole cell but also in the cytoplasmic fraction; capped and polyadenylated ORF0 mRNAs were present in the cytoplasm (Figure 3A).

Most intronic L1s are in the reverse orientation with respect to their host genes (Smit, 1999), including L1s with intact ORF0: ~650 protein coding genes in human and ~450 in chimp contain ORF0 loci in the same direction as host gene transcription (data not shown), raising the possibility of a number of ORF0-host gene fusion events. The sequence logos of ORF0 loci in human and chimp that have the potential to splice, along with commonly used SD sites, are shown in Figure 3B.

To identify ORF0 fusion transcripts in human and chimp, we turned to RNA sequencing (RNA-seq) data that we had generated from iPS cells (Marchetto et al., 2013). Indeed, RNAs for a number of fusion events were observed in both species (Figures 3C, 3D, and S3B). Analysis of the contribution of ORF0 isoforms (iso-0) to the expression of these genes suggested that some genes were primarily driven by ORF0 in iPS cells, whereas for other genes, iso-0 contribution ranged from moderate to minor (Figures 3C and 3D; Table S2). We also extended our analysis to fibroblast-iPS pairs and observed that ORF0 transcript levels were dramatically upregulated in both human and chimp pluripotent stem cells compared to respective source fibroblasts (Figures S3C and S3D and data not shown).

ORF0 mRNAs Are Associated with Ribosomes

The presence of capped ORF0 mRNAs with a polyA tail in the cytoplasm as well as fusion transcripts with proximal exons of protein coding genes prompted us to investigate, by analyzing ribosome footprinting data, whether ORF0 RNAs were associated with ribosomes (ribosome footprinting [Ribo-seq]) (Ingolia et al., 2011). First, we mapped Ribo-seq reads obtained from HEK293T cell line (Shalgi et al., 2013) to L1HS consensus sequence (Figure 3E). In the sense orientation, a plateau of ribosome footprints was detected for ORF1 but ORF2 signal was much weaker, a finding that is in accordance with the known translation levels of ORF1 and ORF2 proteins (Alisch et al.,

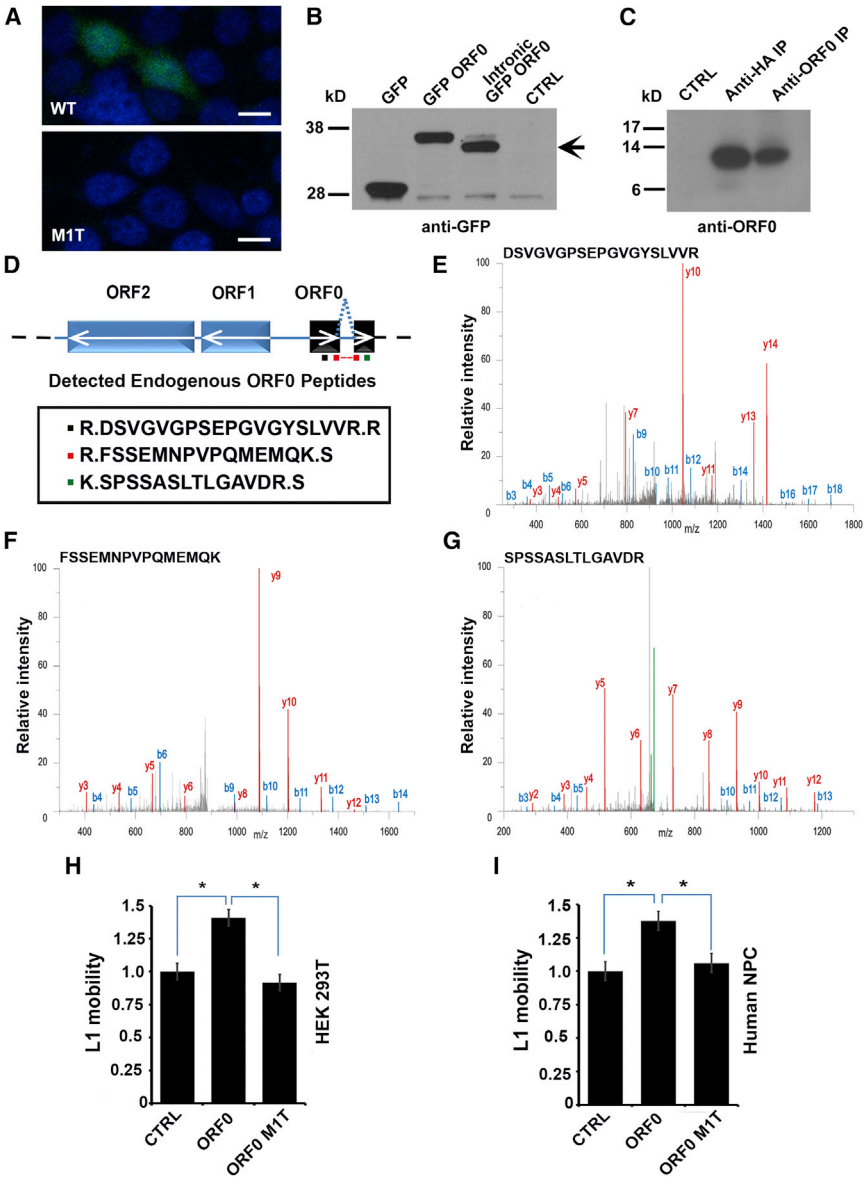


Figure 4. ORF0 Protein: Intronic Expression, Endogenous Detection, and Effect on L1 Mobility

(A) GFP-ORF0 can be expressed from an intronic position in an ORF0 initiator Met-dependent manner. The GFP-ORF0-L1 cassette (wild-type or M1T) was cloned in the antisense orientation in an intron. GFP was detected by confocal microscopy. Scale bar, 10 μm.

(B) Western blot analysis of GFP-ORF0 expression suggests that intronic ORF0 protein is produced but is not full-length. The fusion protein expressed from the intronic construct is indicated with the black arrow.

(C) Functionality of ORF0 antibody was tested using overexpressed protein, by immunoprecipitation, and subsequent western blotting.

(D) Schematic description and sequences of identified ORF0 peptides. The first peptide (black square) resides upstream of SD2. The second peptide (red square) spans the splice junction of proteins formed through splicing between SD2 and SA1. The third peptide (green square) is located downstream of SA1 within the L1 sequence.

(E–G) Spectra of peptides (#1, #2, and #3) identified by proteomic searches. Green peaks in (G) represent neutral losses.

(H and I) Overexpression of ORF0 protein, but not ORF0 RNA, increases L1 mobility based on luciferase L1 reporter in HEK293T cells and human NPCs. Potential antisense RNA effects were controlled for by using a single-nucleotide mutant ORF0 that replaces the initiator Methionine with Threonine. Data are presented as mean ± SEM. *Denotes $p < 0.05$ significance between indicated groups using t test.

See also Figure S4.

2006). In the antisense orientation, a strong signal was evident for ORF0 (Figure 3E). Interestingly, this signal also extended beyond the FL ORF0 sequence, which may be due to within-L1 splicing events (see below and data not shown) or L1s from older families, in which the encoded consensus ORF0 extends until the end of L1 (see Figure S2C). Even though reads obtained by ribosome footprinting were shorter than those gained from RNA-seq, we observed spliced ORF0 footprints of in-frame fusions to *SCAMP1*, *SLC44A5*, *GJB4*, *HTR2C*, and *RABGAP1L* (driven by a human-specific L1 insertion). Thus, the influence of ORF0 may not necessarily be limited to L1 biology.

ORF0-Downstream Exon Fusion Protein Is Expressed

To test whether ORF0 could be transcribed and translated from an intronic position, we cloned the GFP-ORF0-L1 cassette in the antisense orientation within a natural human intron. Upon transfection of this construct into cells, GFP-ORF0 was expressed. Moreover, translation started at the ORF0 1st Met, as the M1T mutation abolished expression (Figure 4A). Interestingly, GFP signal was localized throughout the cell instead of in nuclear foci. This difference in localization was explained by western blot analysis, which showed that intronic GFP-ORF0 fusion protein was different from GFP alone or GFP-FL ORF0, suggesting that a spliced product was translated (Figure 4B). Generation of a fusion protein via splicing between SD1 of ORF0 and the downstream exon was confirmed by sequencing (data not shown).

Proteomic Detection of Endogenous ORF0 Peptides

Having observed ORF0 transcripts as well as expression from reporter plasmids, we investigated endogenous ORF0 products. Proteomic identification of ORF0 requires detection of peptides within unspliced ORF0 or N terminus ORF0 fragments of fusion proteins. Therefore, due to the small size of ORF0, a limited number of possible peptides are available for detection by mass

spectrometry. In addition, the distribution of the target residues (K and R) for trypsin, the most commonly used enzyme for proteomics, leads to the generation of non-ideal peptide fragment sizes (see Figure 2F): the N terminus is poor in these residues whereas the C terminus is rich, generating a very small number of peptides optimal for mass spectrometry. In fact, only one peptide from the main body of ORF0 could be detected in our mass spectrometry analysis of overexpressed ORF0 (Figure S4A).

Nevertheless, we proceeded to attempt detection of endogenous ORF0 peptides. We started by raising polyclonal anti-ORF0 antibodies targeting the consensus L1HS FL-ORF0 protein. Upon confirmation that the ORF0 antibody worked for immunoaffinity enrichment from overexpressed HA-ORF0 extracts (Figure 4C), we turned to the cultured cell type that expressed the highest levels of ORF0 transcripts as a class: human pluripotent stem cells. In parallel, we computationally generated an RNA expression-based ORF0 proteomics database that included potential unspliced and spliced ORF0 proteins. The combined ORF0-Human Uniprot database was used in spectra searches. Next, immunoprecipitates from control and ORF0 antibody were subjected to mass spectrometry analysis. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) spectra searches did not find any ORF0 fragments in control antibody samples. However, searches of anti-ORF0 immunoprecipitates led to identification of endogenous ORF0 peptides (Figures 4D–4G). Spectra obtained from overexpressed peptides for comparison and further information on all the spectra are presented in Figures S4A–S4D. The first peptide (black square) resides upstream of SD2. The second peptide (red square) spans the splice junction of proteins formed through splicing between SD2 and SA1 (SA1: based on RNA-seq analysis, a functional splice acceptor site 336 nucleotides downstream of the ORF0 start site in the L1 5′UTR antisense). The third peptide (green square) is located downstream of SA1 within the LINE-1 sequence (Figure 4D). There are multiple loci that can encode the observed ORF0 peptides and the exact identities of source loci are currently unknown.

ORF0 Enhances L1 Mobility

Given the fact that the ORF0 coding sequence resides in the L1 5′UTR with bidirectional promoter activity, the most parsimonious function for ORF0 would be a potential effect on L1 mobility. Human L1s driven by CMV or CAG promoters are mobile (Moran et al., 1996); thus it is clear that ORF0 is not essential for L1 activity. We attempted to test potential in *cis* effects of ORF0 mutations; however, this task was hampered by the fact that the ORF0 sequence overlaps with the forward L1 promoter (data not shown). Thus, we overexpressed ORF0 in *trans* and tested for its effect on L1 mobility. To prevent any direct antisense L1 RNA effect due to transcription of ORF0, we used a CAG promoter-driven L1 reporter. In HEK293T cells, ORF0 expression led to a ~41% increase in L1 mobility (Figure 4H). To rule out any indirect effects of expressing antisense L1 RNA, we also used the single nucleotide mutant control, ORF0 M1T, that did not produce ORF0 protein. This construct had no effect on L1 mobility, strongly suggesting that ORF0 protein was responsible for the observed increase (Figure 4H). Importantly, wild-type, but not M1T mutant, ORF0 also increased L1 mobility in human embryonic stem (ES) cell-derived neural progenitors (human NPC) by ~38% (Figure 4I), bringing forth the possibility that ORF0 may contribute to somatic variation by enhancing L1 activity in pluripotent cells.

DISCUSSION

The constant competition between transposable elements and host-protective mechanisms contributes to genome evolution (Daugherty and Malik, 2012; Slotkin and Martienssen, 2007). It is currently unclear whether L1 antisense promoter activity has been a major factor in this arms race. From an L1 perspective, antisense transcription can positively influence sense expression through recruitment of transcriptional machinery, inducing open chromatin structure or via formation of a non-coding RNA. On the other hand, expression of antisense RNA can lead to dsRNA formation, which may trigger an RNAi response (Mätlik et al., 2006; Yang and Kazazian, 2006). Our results suggest that, in addition to the aforementioned roles, L1 5′UTR has the ability to initiate translation in the antisense direction.

ORF0 is present in more than ~2,500 loci in the ape genomes, whereas this number is much smaller in the Old World monkeys. While some of this difference may be related to variable genome sequence quality, we expect this difference to mostly represent L1 biology. The alignment of ORF0 sequences from human L1PA1–8 suggests that the main difference between these families is the C terminus of ORF0. We have also noticed that the sequences around the ORF0 translation start site influence forward promoter activity. It is possible that the translation activity in the antisense L1 5′UTR is coupled with the forward promoter activity, and thus the N terminus is more conserved with respect to the rest of the ORF0 sequence due to evolutionary pressure. If that indeed is the case, translation activity in rhesus and baboon may generate distant relatives of the ape ORF0. Consistent with this hypothesis, searches for ORF0 in New World monkey genomes reveal a very small number of loci that have homology to human ORF0, with similarity only at the N terminus. Considering the fact that L1 retrotransposons recruit new 5′UTRs over time, it is conceivable that distant primates such as marmosets and squirrel monkeys may have significantly different 5′UTRs. Improved primate genome sequence quality and future experimentation will allow the testing of these possibilities.

Expression of ORF0, but not an untranslated point mutant version, enhances L1 activity from L1 luciferase mobility reporter in human cells, suggesting a role for ORF0 protein in L1 activity. We currently do not know the mechanism of this effect. Similar to ORF1, ORF0 does not share any extensive homology with known genes, so it is not possible to propose a domain-based prediction. However, ORF0 is a highly positively charged protein that may act by binding to nucleic acids. The PML proximity to ORF0 is intriguing, especially given that a large number of proteins are recruited to PML bodies depending on the cellular state, with stress playing a prominent role in determining the content as well as the morphology of PML bodies. Interestingly, PML is involved in antiviral responses and protects cells from viral infections. Some viral proteins target the integrity of PML bodies and a large number of components are transcriptionally regulated by the interferon pathway (Everett and Chelbi-Alix, 2007). Whether

localization adjacent to PMLs is reflective of ORF0 function or the cell's response remains to be seen. It is possible that ORF0, analogous to some viral proteins, may interfere with the functions of PML and enhance mobility. Further studies will be required to gain insight into the mechanism of action of ORF0.

The influence of ORF0 may not necessarily be limited to L1 biology. Our transcriptomic analysis suggests that exons of host genes provide splice acceptor sites for intronic or proximal ORF0 loci. Overall, ORF0 expression levels correlate with the pluripotency of the cell types and ORF0-proximal exon fusion products are detected by proteomics. While any effects of ORF0 expression on the host or proximal gene would be context and sequence dependent, one could make certain predictions. If the downstream exon is in frame with respect to ORF0 upon splicing, the N terminus of the host protein would be replaced by an ORF0 variant, which could alter the localization and/or function. Out-of-frame ORF0 fusions would contain amino acids from an alternative frame of the gene and most would encounter a stop codon. Such transcripts, depending on the context, might be expressed or be subject to nonsense-mediated decay (NMD). By virtue of high copy numbers and sequence variants, one would expect to see varying degrees of NMD response. In addition, cell-state transitions, stress, and crosstalk with the RNAi pathway might provide opportunities for NMD targets to be translated (Kervestin and Jacobson, 2012). In cases of fusions of ORF0 located upstream of coding sequences, ORF0 might act as an upstream ORF (uORF). Since uORF function is affected by the length and sequence of the uORF as well as by the distance between the upstream and the main ORF, variations in ORF0 sequences could result in differential translation regulation (Andrews and Rothnagel, 2014).

L1s, as the sole autonomously active retrotransposons in primate genomes, continue to shape our genomes. Our data suggest that, in addition to their previously ascribed roles in gene regulation (Huang et al., 2012), L1s contain a third ORF and have the ability to generate insertion site-dependent ORFs via splicing. Considering the fact that transcription and translation start within L1 elements, these ORF0 variants could be co-regulated. Analogous to the other L1 proteins, disorders such as neoplasms (Rodić et al., 2014) may provide opportunities for higher ORF0 expression, which in turn could contribute to the pathological phenotypes. It is tempting to speculate that, over evolutionary time, the propensity of ORF0 to splice into proximal exons may have led to not only gene regulatory changes but also the emergence of new proteins. The extent to which ORF0 variants contribute to diversity, both in evolutionary terms and disease conditions, remains to be investigated.

EXPERIMENTAL PROCEDURES

Cloning and Mutagenesis

Primers from IDT and Phusion High Fidelity Polymerase (NEB) were used for PCRs. pGL4.10 (Promega) was the plasmid backbone used for promoter luciferase assays. To test the effect of ORF0 expression on L1 mobility, ORF0 promoter, coding sequence, and the downstream sequence (until the end of L1 in the antisense orientation) was cloned into pEF-BOS-EX (Mizushima and Nagata, 1990). To include any potential within-L1 splicing products and prevent contribution from the plasmid backbone, a fragment containing stop codons in all three frames as well as a polyA signal was included immediately downstream of the insert. ORF0-GFP construct was cloned into a modified (SV40 promoter and luciferase removed) pSICheck2 vector (Promega). A modified (luciferase cassette removed) pYX014 plasmid (Xie et al., 2011) was used for GFP-ORF0 and mCherry-ORF0 cloning: nucleotide 13 of ORF0 was mutated (C → G) to generate an AscI site that was used for subsequent cloning of GFP and mCherry. HA-tagged ORF0 was cloned into pCDNA3.1 for in vitro translation. GFP-ORF0-L1 cassette was cloned into pEF-BOS-EX with BglII for intronic expression. Mutagenesis was carried out using the Quick Change II XL Site Directed Mutagenesis Kit (Agilent Technologies).

RNA Extraction, Reverse Transcription, and cDNA Preparation

RNA was prepared using Trizol (Invitrogen). cDNA was synthesized using the Superscript III First Strand Synthesis System for RT-PCR (Invitrogen).

Cell Culture and Transfection

HEK293T cells (ATCC) were cultured in DMEM$^+$ GlutaMax medium (Life Technologies) supplemented with 10% fetal bovine serum (Omega Scientific) and grown at 37°C in 5% CO_2. Cells were transfected using polyethylenimine (Polysciences). HUES6 human ES cells were cultured feeder-free on Matrigel-coated dishes (BD) using mTeSRTM1 (StemCell Technologies) and passaged once every 3–4 days using Collagenase type IV enzyme.

Human NPC Derivation, Growth, and Nucleofection

NPCs were differentiated from HUES6 cells through embryoid body and rosette generation and grown as previously described (Marchetto et al., 2010). Plasmid delivery into human NPCs was performed by nucleofection (Lonza/Amaxa Nucleofector, kit VPG-1005).

In Vitro Translation

ORF0 was synthesized in vitro by employing the TNT Coupled Reticulocyte Lysate System (Promega) using T7 polymerase.

Cell Extracts and Western Blot Analysis

Cells were harvested 2 days post transfection, washed with cold DPBS, and lysates were prepared with ice cold RIPA lysis buffer (50 mM Tris-HCl [pH 7.4], 150 mM NaCl, 0.25% deoxycholic acid, 1% NP-40, 0.1% SDS, and 1 mM EDTA) containing complete protease inhibitor cocktail with EDTA (Roche) and 1 mM DTT. Lysates were incubated on ice for 15 min, spun at 14,000 × *g* for 15 min at 4°C, and the supernatants were collected. Primary antibodies: rabbit α-GFP (1:2000, Santa Cruz sc-8334), rat α-HA peroxidase high-affinity 3F10 (1:1000, Roche), and α-ORF0 (1:300). Secondary antibody: (1:5,000, GE NA934).

Fluorescence Detection

Cells were grown in poly-L-lysine (Sigma) coated 2-well LabTek chamber slides (Nunc, Fisher), fixed in 4% paraformaldehyde (Sigma) for 15 min at room temperature, and washed with TBS. The nuclei were stained with DAPI (1:1,000, Sigma) and the slides were mounted using polyvinyl alcohol with DABCO (Sigma).

Computational Analyses

Detection and visualization of ORF0 loci in human and chimp genomes: the UCSC genome browser and Ensembl databases were used to retrieve potential ORF0 coding sequences, which were subsequently in silico translated. The Ensembl databases (hg19, panTro4) were used for blastn, allowing some local mismatch but no gap to obtain ORF0 loci. An alternative method of retrieving all potential full-length ORF0 sequences from RepeatMasker was tested and led to very similar results. Custom python scripts and EMBOSS suite (Rice et al., 2000) were used for identification and characterization of ORF0 loci, full-length as well as untruncated-until-splice-donor, in the genome. Sequences that did not contain a GT dinucleotide at the splice donor site were removed. Ensembl Karyotype View tool was used for visualization of the ORF0 loci. Upon confirmation of an annotation error in the Chimp Chr 2B, the erroneous fragment was removed from the image. The removed region contained no genes or TEs. Analysis of RNA-seq datasets: RNA-seq (human and chimp iPS cells) data from GEO: GSE47626 (Marchetto et al., 2013), GEO: GSE44646 (Wang et al., 2014), GEO: GSE60996 (Gallego Romero et al., 2015), and ArrayExpress: E-MTAB-2031 (Chan et al., 2013); CAGE

(capped 5′ RNA-seq) data from GEO: GSE34448 (Djebali et al., 2012); Ribo-seq (ribosome footprinting) data from GEO: GSE32060 (Shalgi et al., 2013) were analyzed from raw FASTQ files in a consistent manner. Reads were aligned to the reference human (hg19) and chimpanzee (panTro4) genomes with STAR, which is capable of identifying novel splice junctions (Dobin et al., 2013). Spliced ORF0 reads were identified by filtering out all multimappers and only considering reads originating from an ORF0 locus (direct overlap of 5′ end for stranded RNA-seq and direct overlap of either read end for unstranded RNA-seq). Read distributions along L1 were found by aligning reads to the consensus L1HS element using STAR (Dobin et al., 2013). Read densities along the + and – strands were further normalized based on the total number of reads in each experiment that were alignable to the full genome. Ratios of isoforms (ORF0 versus total) were determined by comparing the splice junction reads (j) of ORF0, j(0c), to j(ab), j(bc), j(cd), j(de), where the order of exons are a-b-0-c-d-e: ratio = average((j(0c)/(average(j(ab),j(bc)) + j(0c))), (j(0c)/average(j(cd) + j(de)))). This allowed us to get a more reliable estimate compared to calculations that rely solely on ratio at one exon j(0c) and j(bc) and to reduce the 3′ bias that is observed in polyA-based sequencing. In the few cases where the ratio is higher than 1 (maximum being 1.2), these ratios are presented as 1 in the tables (Figures 3C, 3D, and Table S2). Genes in the tables went through further manual inspection. Proteomic database generation: RNA-seq reads from human iPS/ES cells were assembled using Cufflinks, ORF0 containing transcripts were selected and redundancies were removed. In parallel, ORF0-containing mRNAs that are either ESTs or annotated transcripts were added to the RNA-seq list. The combined list was in silico translated and appended to the current human Uniprot database for spectra searches. Determination of species that have ORF0 loci: L1HS/L1Pt consensus ORF0 sequence (identical) was used in Blat and blast searches to determine the genomes that contain ORF0 loci. The absence of ORF0 loci in non-Catarrhine primates was further confirmed by in silico translation of L1 sequences (with repeat start <1,000) and subsequent search for loci that can encode a polypeptide with > = 50% identity to ORF0 protein (FL or SD1-ORF0). Generation of consensus primate ORF0 sequences for phylogenetic analysis: ORF0 loci that can encode an untruncated protein >210 nucleotides were retrieved via blast searches and subsequent in silico translation and filtering. The sequences were trimmed to 213 nucleotides (length of FL ORF0) and used in molecular phylogenetic analysis. Basic molecular phylogenetic analysis: Clustal Omega was used to generate the alignments. The evolutionary history was inferred by using the maximum likelihood (ML) method based on the JTT matrix-based model. The tree with the highest log likelihood is shown. A total of 1,000 bootstrap replicates were used for test of phylogeny. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model and then selecting the topology with superior log likelihood value. The tree is drawn to scale, and branch lengths represent number of substitutions per site. The analysis involved amino acid sequences and all positions with <95% site coverage were eliminated. Muscle generated alignments as well as maximum parsimony analysis generated a very similar tree. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013). Analyses using RAxML and PhyML as well as neighbor joining methods resulted in very similar trees. DNA and protein logos were generated using WebLogo (Crooks et al., 2004).

L1 Mobility Assays

Luciferase-based L1 mobility reporters used were previously described (Xie et al., 2011). Cells were transfected/nucleofected with experimental constructs together with L1 mobility reporter plasmid pYX017. Luciferase activity was quantified at day 3 using the Dual Luciferase Reporter 1000 Assay System (Promega, E1980) and a Perkin Elmer Victor X Luminometer. A two-tailed t test was used for statistical analysis.

Promoter Activity Assays

Promoter activity was measured by co-transfecting ORF0 promoter constructs cloned into pGL4.10 (Promega) along with the normalization vector phRLTK (Promega). Activity was measured after 2 days, as in the L1 activity assays. A two-tailed t test was used for statistical analysis.

Antibody Generation and Immunoprecipitations

Peptides corresponding to ORF0 amino acid residues 20-34, 33-49 and 50-65 in the L1HS consensus were synthesized, conjugated to KLH and used in generation of rabbit polyclonal antibodies (Covance). For immunoprecipitations (IPs), cells were washed with DPBS, collected, and frozen. Cell pellets were thawed in mDm lysis buffer (25 mM Tris [pH 7.5], 150 mM NaCl, 1.5 mM $MgCl_2$, 1% Triton X-100, 1 mM DTT, protease inhibitors [Roche]) and supernatant from a 15,000 × *g* 15 min spin was used in IPs. Control and ORF0 antibodies were conjugated to magnetic beads (Pierce). IP duration was 4–6 hr, washes were done with the mDm buffer, and beads were heated to 95°C for 10–12 min for elution.

Proteomic Sample Prep and Analysis

Samples were precipitated by methanol/chloroform. Dried pellets were dissolved in 8 M urea/100 mM Tris, [pH 8.5]. Proteins were reduced with 5 mM tris(2-carboxyethyl) phosphine hydrochloride (TCEP, Sigma-Aldrich) and alkylated with 10 mM iodoacetamide (Sigma-Aldrich). Proteins were digested overnight at 37°C in 2 M urea/100 mM Tris, [pH 8.5], with trypsin (Promega). Digestion was quenched with formic acid, 5% final concentration and a final volume of 50 μl.

The digested samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo). Samples were analyzed with injections of 8 μl of the protein digest per LC/MS run. The digest was injected directly onto a 40-cm, 75-μm ID column packed with BEH 1.7 μm C18 resin (Waters). Samples were separated at a flow rate of 200 nl/min on a nLC 1000 (Thermo). Buffer A and B were 0.1% formic acid in water and acetonitrile, respectively. Two reverse phase gradients of 140 min and 450 min were used to maximize sampling efficiency of the digest. Ninety percent buffer B was used for 10 min final washes at the ends of gradients. Column was re-equilibrated with 20 μl of buffer A prior to the injection of sample. Peptides were eluted directly from the tip of the column and nanosprayed into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap Fusion was operated in a data-dependent mode. Full MS^1 scans were collected in the Orbitrap at 120 K resolution with a mass range of 400–1,600 m/z and an AGC target of $5e^5$. The cycle time was set to 3 s, and within this 3 s the most abundant ions per scan were selected for CID MS/MS in the ion trap with an AGC target of $1e^4$ and minimum intensity of 5,000. Maximum fill times were set to 50 ms for MS scans and 100 and 35 ms for MS/MS scans in the 140 min and 450 min methods, respectively. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled and dynamic exclusion was used with an exclusion duration of 5 s.

Protein and peptide identification were done with Integrated Proteomics Pipeline- IP2 (Integrated Proteomics Applications). Tandem mass spectra were extracted from raw files using RawConverter and searched with ProLuCID against ORF0-human UniProt database. The search space included all fully tryptic and half-tryptic peptide candidates. Carbamidomethylation on cysteine was considered as a static modification. Data were searched with 50 ppm precursor ion tolerance and 500 ppm fragment ion tolerance. Data were filtered to 10 ppm precursor ion tolerance post search. Identified proteins were filtered using DTASelect (Tabb et al., 2002) and utilizing a target-decoy database search strategy to control the false discovery rate to 1% at the protein level.

Imaging

All imaging was carried out using a Zeiss LSM 780 Confocal Microscope. Images were taken using either a 20× or a 100× oil objective. The z stack intervals were 1 μm. Image analysis was performed with ZEN (Zeiss) and Imaris (Bitplane). Both PML and ORF0 foci were identified using the Spots object on Imaris (Bitplane) using a fixed spot size of 0.5 μm (the measured average XY diameter of nuclear bodies).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, two tables, and one movie and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.025.

AUTHOR CONTRIBUTIONS

A.M.D. is the lead author, designed the study, and was involved in execution of all wet lab, computational studies, and data analysis. I.N. contributed to the concept and performed in vitro translation assays. B.E.K. contributed to the concept, performed imaging, and image processing. M.P. provided technical support in performing all wet lab experiments (including imaging) and helped with manuscript preparation. C.B. contributed to the concept, provided bioinformatics guidance, and performed the initial computational analysis of RNA-seq datasets. M.C.N.M. performed experiments involving human NPC derivation, culture, and nucleofection. J.K.D. and J.J.M. performed proteomic sample prep and analysis. J.K.D., J.J.M., A.A., and J.M. performed proteomic searches. L.M. provided technical support. A.S. contributed to the concept and provided proteomic guidance. F.H.G. is the senior author, contributed to the concept, analyzed the data, revised the manuscript, and provided financial support. A.M.D. and F.H.G. wrote the manuscript. All authors contributed comments on the manuscript.

ACKNOWLEDGMENTS

This work was supported by funds from the Leona M. and Harry B. Helmsley Charitable Trust, the JPB Foundation, and the G. Harold and Leila Y. Mathers Charitable Foundation. We thank Gökhan Şentürk, Ruth Keithley, Ana Mendes, Dilara Halim, and Iryna Gallina for technical help; Sara Linker, Christos Tzitzilonis, and Stephane Boissinot for discussions; Peter Hemmerich and Wenfeng An for reagents; James Fitzpatrick and Michael Adams for help with imaging; Mary Lynn Gage for editorial comments on the manuscript; and everybody involved in GEO and ENA database generation and data contributions.

Received: January 4, 2015
Revised: July 7, 2015
Accepted: August 25, 2015
Published: October 22, 2015

REFERENCES

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J.V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. *20*, 210–224.

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.

Bernardi, R., and Pandolfi, P.P. (2007). Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. Nat. Rev. Mol. Cell Biol. *8*, 1006–1016.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA *100*, 5280–5285.

Chan, Y.S., Göke, J., Ng, J.H., Lu, X., Gonzales, K.A., Tan, C.P., Tng, W.Q., Hong, Z.Z., Lim, Y.S., and Ng, H.H. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663–675.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Cruickshanks, H.A., and Tufarelli, C. (2009). Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics *94*, 397–406.

Cruickshanks, H.A., Vafadar-Isfahani, N., Dunican, D.S., Lee, A., Sproul, D., Lund, J.N., Meehan, R.R., and Tufarelli, C. (2013). Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer. Nucleic Acids Res. *41*, 6857–6869.

Daugherty, M.D., and Malik, H.S. (2012). Rules of engagement: molecular insights from host-virus arms races. Annu. Rev. Genet. *46*, 677–700.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805–1808.

Everett, R.D., and Chelbi-Alix, M.K. (2007). PML and PML nuclear bodies: implications in antiviral defence. Biochimie *89*, 819–830.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. *41*, 563–571.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell *87*, 905–916.

Gallego Romero, I., Pavlovic, B.J., Hernando-Herraez, I., Zhou, X., Ward, M.C., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., Mitrano, A., et al. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. eLife *4*, e07103.

Goodier, J.L., Zhang, L., Vetter, M.R., and Kazazian, H.H., Jr. (2007). LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. Mol. Cell. Biol. *27*, 6469–6483.

Hancks, D.C., and Kazazian, H.H., Jr. (2012). Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. *22*, 191–203.

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., and Kazazian, H.H., Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum. Mol. Genet. *20*, 3386–3400.

Huang, C.R., Burns, K.H., and Boeke, J.D. (2012). Active transposition in genomes. Annu. Rev. Genet. *46*, 651–675.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Kervestin, S., and Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. Nat. Rev. Mol. Cell Biol. *13*, 700–712.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. *16*, 78–87.

Kleckner, N. (1990). Regulation of transposition in bacteria. Annu. Rev. Cell Biol. *6*, 297–327.

Kolosha, V.O., and Martin, S.L. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc. Natl. Acad. Sci. USA *94*, 10155–10160.

Kozak, M. (1987). An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. *15*, 8125–8148.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595–605.

Macia, A., Muñoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. Mol. Cell. Biol. *31*, 300–316.

Marchetto, M.C., Carromeu, C., Acab, A., Yu, D., Yeo, G.W., Mu, Y., Chen, G., Gage, F.H., and Muotri, A.R. (2010). A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. Cell *143*, 527–539.

Marchetto, M.C., Narvaiza, I., Denli, A.M., Benner, C., Lazzarini, T.A., Nathanson, J.L., Paquola, A.C., Desai, K.N., Herai, R.H., Weitzman, M.D., et al. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. Nature *503*, 525–529.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D., and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science *254*, 1808–1810.

Mätlik, K., Redik, K., and Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. J. Biomed. Biotechnol. *2006*, 71753.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA *36*, 344–355.

Mizushima, S., and Nagata, S. (1990). pEF-BOS, a powerful mammalian expression vector. Nucleic Acids Res. *18*, 5322.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917–927.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530–1534.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Rodić, N., Sharma, R., Sharma, R., Zampella, J., Dai, L., Taylor, M.S., Hruban, R.H., Iacobuzio-Donahue, C.A., Maitra, A., Torbenson, M.S., et al. (2014). Long interspersed element-1 protein expression is a hallmark of many human cancers. Am. J. Pathol. *184*, 1280–1286.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics *1*, 113–125.

Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013). Widespread regulation of translation by elongation pausing in heat shock. Mol. Cell *49*, 439–452.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA *100*, 15776–15781.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. *9*, 657–663.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell. Biol. *21*, 1973–1985.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol. Cell. Biol. *10*, 6718–6729.

Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd. (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. *30*, 2725–2729.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V., et al. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature *516*, 405–409.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res. *15*, 1073–1078.

Xie, Y., Rosser, J.M., Thompson, T.L., Boeke, J.D., and An, W. (2011). Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. Nucleic Acids Res. *39*, e16.

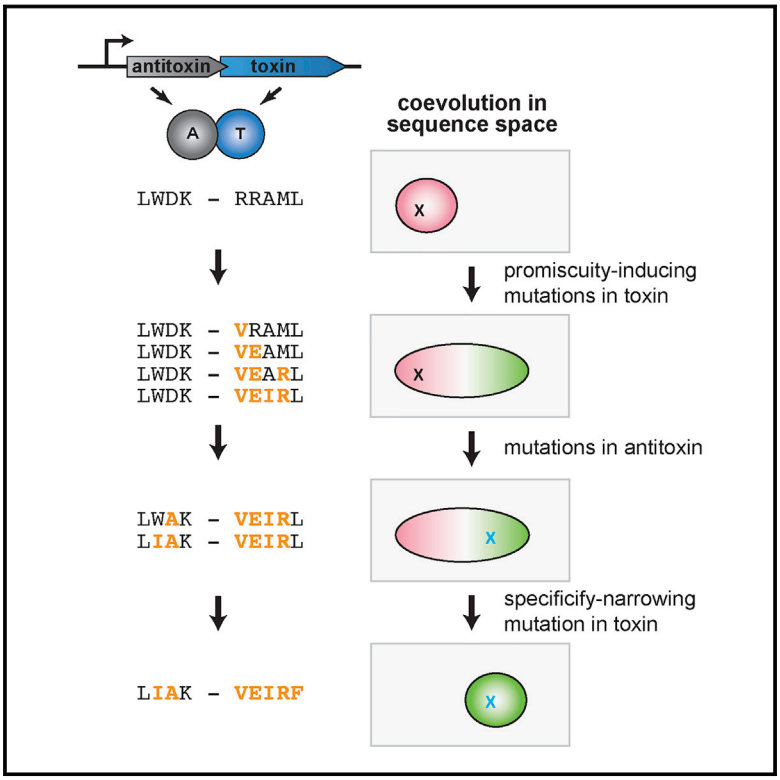
Yang, N., and Kazazian, H.H., Jr. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat. Struct. Mol. Biol. *13*, 763–771.

Article

Cell

Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates

Graphical Abstract



Highlights

- ParD-ParE toxin-antitoxin systems interact in a highly specific manner
- Toxin-antitoxin systems can coevolve without ever disrupting their interaction
- Promiscuous variants can serve as mutational intermediates during coevolution
- Promiscuous variants are abundant in sequence space and connected to specific variants

Authors

Christopher D. Aakre, Julien Herrou, Tuyen N. Phung, Barrett S. Perchuk, Sean Crosson, Michael T. Laub

Correspondence

laub@mit.edu

In Brief

Interacting proteins can coevolve through the generation of promiscuous variants, which serve as mutational intermediates that preserve the ability of the two proteins to functionally interact while they evolve.

Aakre et al., 2015, Cell *163*, 1–13
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.055

CellPress

Please cite this article in press as: Aakre et al., Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.055

Cell

Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates

Christopher D. Aakre,[1] Julien Herrou,[3] Tuyen N. Phung,[1] Barrett S. Perchuk,[1] Sean Crosson,[3] and Michael T. Laub[1,2,*]
[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA
*Correspondence: laub@mit.edu
http://dx.doi.org/10.1016/j.cell.2015.09.055

SUMMARY

Interacting proteins typically coevolve, and the identification of coevolving amino acids can pinpoint residues required for interaction specificity. This approach often assumes that an interface-disrupting mutation in one protein drives selection of a compensatory mutation in its partner during evolution. However, this model requires a non-functional intermediate state prior to the compensatory change. Alternatively, a mutation in one protein could first broaden its specificity, allowing changes in its partner, followed by a specificity-restricting mutation. Using bacterial toxin-antitoxin systems, we demonstrate the plausibility of this second, promiscuity-based model. By screening large libraries of interface mutants, we show that toxins and antitoxins with high specificity are frequently connected in sequence space to more promiscuous variants that can serve as intermediates during a reprogramming of interaction specificity. We propose that the abundance of promiscuous variants promotes the expansion and diversification of toxin-antitoxin systems and other paralogous protein families during evolution.

INTRODUCTION

Many interacting proteins within the same cell, particularly signaling proteins, are members of large paralogous families that have expanded through duplication and divergence. To expand in number, paralogous interacting proteins typically must become specific after duplication to avoid unwanted cross-talk (Capra et al., 2012; Zarrinpar et al., 2003). The specificity determinants of protein-protein interactions remain poorly defined in most systems. Even in the cases where they have been identified, we lack a detailed understanding of how a new, insulated protein-protein interaction emerges during the course of evolution and, more generally, the mutational paths followed during protein evolution (DePristo et al., 2005).

Computational studies demonstrate that interacting proteins often coevolve. Indeed, identification of coevolving residues has helped guide identification of the specificity determinants of many protein-protein interfaces (Ovchinnikov et al., 2014; Skerker et al., 2008). The implicit notion or underlying model behind these analyses is usually that an interaction-disrupting mutation in one protein can be rescued by a mutation in its partner (Figure 1A). This model, which we call the compensatory mutation model, implies that the system passes through a non-functional or non-interacting state. However, such a state is highly unlikely, particularly for a protein-protein interaction that is critical for the viability of an organism. Alternatively, the specificity of a given protein-protein interaction could change, and become insulated from other paralogous systems, if one of the proteins passes through a promiscuous intermediate (Figure 1B). In this model, an initial mutation in protein A would broaden its specificity, enabling its partner, protein B, to accumulate a mutation that would have disrupted its interaction with the original, ancestral form of protein A. A subsequent mutation in protein A would then narrow its specificity to include the derived, but not the ancestral, form of protein B. In this promiscuous intermediate model, the specificities of the interacting proteins change without ever transitioning through a non-functional intermediate state. Note that in both models, A and B continue to interact through the same set of interfacial residues and do not evolve an alternative interface de novo (Kuriyan and Eisenberg, 2007).

Which of the two models in Figure 1 applies to most pairs of interacting proteins is unclear. In each case, the mutational trajectory involved would produce a signature of pairwise amino-acid coevolution in the phylogenetic record. However, only the latter, promiscuous intermediate model invokes the existence of mutations that are transiently introduced to broaden the specificity of one of the two proteins. The prevalence of such promiscuous states is unknown, as is whether they are easily reached from more specific, extant states.

Bacterial toxin-antitoxin (TA) systems provide an excellent model system for dissecting the coevolutionary dynamics of protein-protein interactions. Originally identified on plasmids, these systems are widely found in bacterial chromosomes, with many species encoding multiple, paralogous copies that share extensive similarity at the sequence and structural levels (Leplae et al., 2011). The biological function of TA systems is unclear, but they have been implicated in stress responses, resistance to phage, formation of persister cells, and bacterial pathogenicity (Yamaguchi et al., 2011). Typically, the toxin is a stable, globular protein that can inhibit cell growth or viability unless antagonized by a cognate antitoxin that directly binds and sequesters the toxin. Changes in the degradation rate or synthesis of the antitoxin can trigger release of the toxin. A toxin is typically encoded in

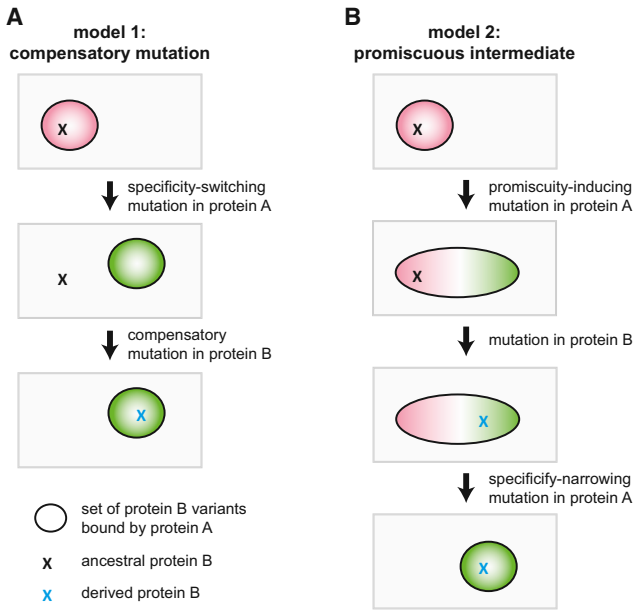


Figure 1. Models for the Evolution of New Protein-Protein Interaction Specificity

(A) In a model of coevolution through compensatory mutations, an initial mutation in protein A that disrupts the A-B interaction is rescued by a compensatory mutation in protein B. Ovals represent the set of protein B variants that are bound by protein A, and Xs indicate particular protein B variants. Note that the intermediate state is a non-functional interaction.

(B) In an alternative model for protein coevolution, protein A first accumulates a mutation that broadens its specificity, followed by a second mutation in protein B that retains its interaction with the new form of A but that would have disrupted its interaction with the ancestral form of protein A. In a final step, protein A mutates to narrow its specificity to include the derived, and not ancestral, form of protein B.

the same operon as an antitoxin, and toxin-antitoxin paralogs frequently arise through operon duplications. An unresolved question is whether toxin-antitoxin systems interact in an exclusive one-to-one manner. Genetic data suggest that these interactions may be specific (Fiebig et al., 2010), and the growth inhibitory effects of a toxin are usually rescued only by expressing its co-operonic antitoxin (Hallez et al., 2010; Ramage et al., 2009). However, interaction specificity has only been directly tested in a limited number of cases, and some groups have suggested that toxins and antitoxins encoded in different operons are capable of interacting in vivo and in vitro, possibly forming large, promiscuous networks (Yang et al., 2010; Zhu et al., 2010).

Here, we systematically measure the binding preferences of 20 ParD-ParE TA family members and find that these toxins and antitoxins are highly specific, interacting almost exclusively with their partner from the same operon. This specificity is encoded by a small set of coevolving residues at the toxin-antitoxin interface, and mutations in these residues are sufficient to reprogram a ParD antitoxin to interact with non-cognate ParE toxins. Guided by these findings, we generated a library with $\sim 10^4$ variants of the key, specificity-determining residues in a ParD antitoxin and selected mutants that antagonize the cognate toxin, a non-cognate toxin, or both. Strikingly, we find that promiscuous variants that antagonize multiple toxins are easily obtained and are also highly connected in sequence space to specific variants. These results suggest that mutational paths leading to changes in toxin-antitoxin specificity are likely to involve promiscuous intermediates. Such paths enable the reprogramming of toxin-antitoxin specificity through the pairwise coevolution of interfacial residues, but without passing through an intermediate state that disrupts the protein-protein interaction. The abundance of promiscuous states likely facilitates the evolutionary expansion of these and other paralogous protein families following operon and whole-genome duplications during evolution.

RESULTS

Toxins and Antitoxins from the ParDE Family Exhibit High Interaction Specificity

To systematically measure the interaction specificity of TA systems, we focused on the ParD-ParE family, which is often found in multiple copies on bacterial chromosomes (Fiebig et al., 2010; Leplae et al., 2011) (Figure S1A). We initially cloned the three chromosomally encoded ParD-ParE pairs from the α-proteobacterium *Mesorhizobium opportunistum* into vectors that allow for separate and inducible expression of the ParE toxin and ParD antitoxin. To measure the interaction specificity for these pairs, we then co-transformed all pairwise combinations of toxin and antitoxin plasmids into *E. coli* and assessed whether the induced expression of each ParD antitoxin rescues the growth arrest resulting from inducing each ParE. As a control, we first confirmed that inducing each ParE toxin inhibited growth of *E. coli* (Figure 2A). Then, plating on a medium that induces both ParD and ParE, we observed growth for each of the three cognate ParD-ParE pairings (Figure 2A). No growth was observed for the six non-cognate pairs, indicating that the ParD antitoxins from *M. opportunistum* can only neutralize their cognate ParE toxins.

We extended this analysis to the 20 chromosomally encoded ParDE pairs from eight different bacteria, including the three pairs from *M. opportunistum* (Figure S1B). For this 20 × 20 matrix of ParD and ParE pairs we observed strong interactions between all 20 co-operonic ParDE pairs, but only 11 of the 380 (or 3%) other possible pairings (Figure 2B). Importantly, these cross-reactions were only observed between ParD and ParE proteins not encoded in the same species, indicating that the ParDE pairs within a given organism are typically insulated from one another. These results indicate that ParD antitoxins are highly specific for their cognate ParE toxins.

Identification of Covarying Residues in ParD and ParE

As a first step in understanding the molecular basis of specificity in ParD-ParE complexes, we solved a 1.59-Å cocrystal structure of the *M. opportunistum* ParD3 antitoxin bound to ParE3, its cognate toxin. This structure revealed a heterotetrameric asymmetric unit composed of ParD3 and ParE3 dimers (Figure S2A), similar to a *C. crescentus* ParD-ParE structure (Dalton and Crosson, 2010). Crystal packing and an estimated mass of ~87 kDa in solution indicate that the biological assembly is composed of two tetramers (Figures S2B and S2C). Within this complex, each ParD3 subunit makes extensive contacts with a

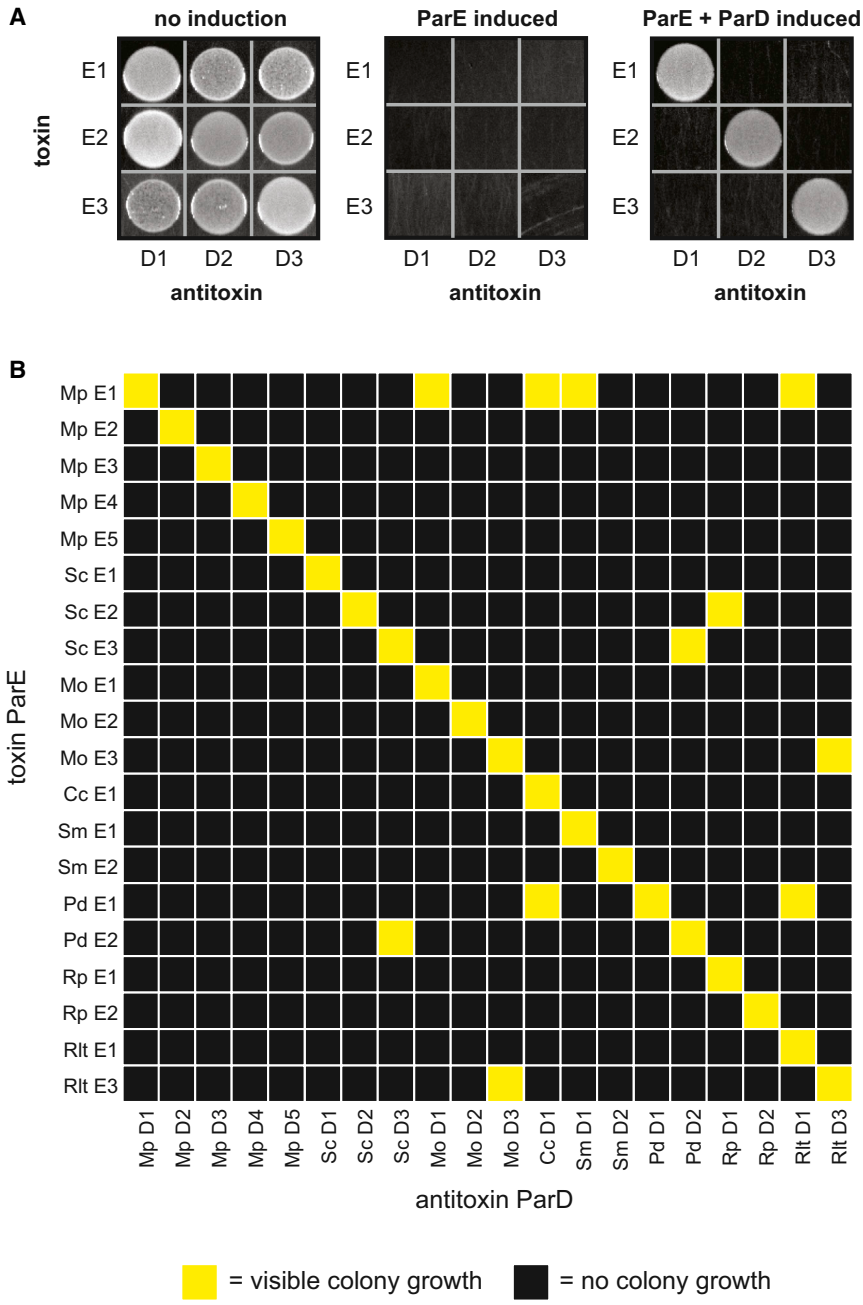


Figure 2. Toxins and Antitoxins from the ParD-ParE Family Exhibit High Interaction Specificity

(A) Testing of interaction specificity for ParD antitoxins and ParE toxins from *Mesorhizobium opportunistum*. Plasmids harboring the toxins and antitoxins indicated were co-transformed into *E. coli* with ParD and ParE induced as indicated.

(B) Comprehensive testing of interaction specificity for 20 ParD and ParE pairs from eight different species. Cells containing each possible ParD-ParE pair were grown on plates that induce the toxin and antitoxin, respectively, and grown overnight at 37°C. Yellow, visible colonies following serial dilution; black, no visible colonies.

See Figure S1.

neighboring ParE3 subunit primarily through its second and third alpha helices, with a total buried surface area of 1,624 Å^2 (Figure 3A).

Previous work with bacterial two-component signaling systems demonstrated that their interaction specificity is controlled by a subset of residues at the protein-protein interface formed by a histidine kinase and response regulator (Skerker et al., 2008). These specificity-determining residues coevolve to maintain the interaction between cognate signaling proteins. Thus, to pinpoint the residues that contribute to the specificity of ParD-ParE interactions, we used GREMLIN, a pseudo-likelihood-based model for coevolution (Kamisetty et al., 2013; Ovchinnikov et al., 2014), to search for residues that strongly covary in a multiple sequence alignment of concatenated, co-operonic ParD and ParE proteins. This analysis identified 10 residues in ParD and 11 residues in ParE that coevolve most strongly. Hereafter, we call these 21 amino acids "specificity" residues, as our work below indicates that they play the dominant role in determining partner specificity. Mapping these specificity residues onto the ParD3-ParE3 crystal structure indicated that they cluster into two groups at the primary molecular interface formed by these proteins (Figures 3B and 3C). The first group sits at the base of the second alpha helix in ParD3 and covaries with residues in the three-stranded beta sheet in ParE3. The second group clusters in the third alpha helix in ParD3 and covaries with residues in the first and second alpha helices of ParE3. We also used GREMLIN to identify residues within each protein (four in ParD and six in ParE) that coevolve with the specificity residues (Figure 3C and S3A). These "supporting" residues may indirectly contribute to ParD-ParE interaction specificity by influencing the orientation or packing of the interfacial specificity residues.

Covarying Residues Dictate Interaction Specificity in the ParD-ParE Family

To determine whether the coevolving residues identified are sufficient to dictate interaction specificity of the ParD-ParE family, we constructed a series of chimeric proteins in which different regions of the *M. opportunistum* ParD3 were replaced with the corresponding regions of ParD1 or ParD2 (Figure S3B). Replacing the entire C-terminal region of ParD3 with the corresponding

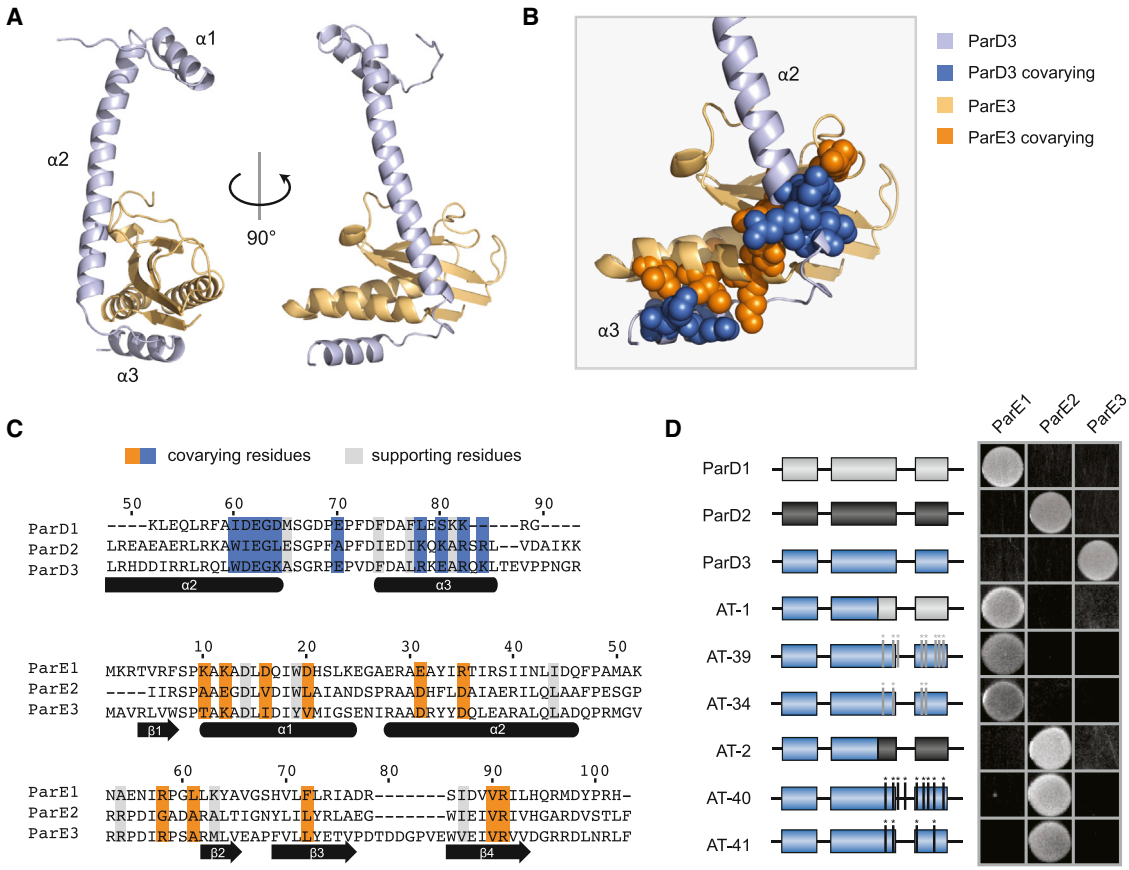


Figure 3. Covarying Residues Dictate Interaction Specificity in the ParD-ParE Family

(A) Structure of the *M. opportunistum* ParD3-ParE3 complex (PDB: 5CEG). Light orange, ParE3 monomer; light blue, ParD3 monomer.

(B) A section of the ParD3-ParE3 structure from (A) magnified; covarying residues shown in space-filling representation.

(C) Alignment of *M. opportunistum* ParD and ParE paralogs with coevolving residues highlighted in blue or orange for ParD or ParE, respectively. Supporting residues, which coevolve with the interfacial coevolving residues, are highlighted in gray.

(D) Mutations in the C terminus of ParD3 can reprogram interaction specificity. The indicated ParD3 mutants were tested against each ParE homolog from *M. opportunistum* using the *E. coli* toxicity-rescue assay.

Also see Figures S2 and S3.

region of ParD1 or ParD2 produced a chimera that lost its ability to interact with ParE3 but gained the ability to interact with ParE1 or ParE2 (Figure 3D). These chimeras involved both clusters of interfacial residues identified as coevolving between ParD and ParE proteins. Replacing only one of these clusters in the ParD3 C terminus was sometimes sufficient to reprogram specificity, but depended on the toxin tested (Figure S3C). These results indicate that the C-terminal region of ParD, which contains the specificity and supporting residues, is sufficient to dictate interaction specificity.

To pinpoint the residues required for interaction specificity, we focused additional mutagenesis on the coevolving residues identified computationally. We generated variants of ParD3 in which all of the specificity and supporting residues were replaced with the corresponding residues in ParD1 or ParD2, for a total of 8 or 9 substitutions, respectively. In each case, we found that these mutations were sufficient to reprogram ParD3 to interact with ParE1 or ParE2 and lose its ability to interact with ParE3 (Figure 3D). Interestingly, ParD3 could be reprogrammed to interact with ParE1 or ParE2 with fewer substitutions. For example, we found sets of four substitutions that were sufficient to reprogram ParD3 to interact with ParE1 or ParE2 (Figure 3D). Taken together, our results indicate that mutating the most highly coevolving residues in an antitoxin can be sufficient to reprogram its interaction specificity, and, in some cases, mutating only a subset of these residues allows a complete switch in partner specificity.

High-Throughput Mapping of Interface Mutant Fitness

The results presented above indicate that antitoxin interaction specificity can be reprogrammed by changing just four residues. But how does specificity change as these four individual substitutions are introduced and does the substitution order matter? Does the specificity of antagonizing one ParE toxin to another change abruptly, or are there promiscuous mutational intermediates? To answer these questions, we sought to generate a large

library of ParD3 variants that included combinations of residues shown to be specific for antagonizing ParE3 or ParE2, as well as the mutational intermediates separating these specific states. To this end, we generated a library of mutants at four of the key interfacial positions in the ParD3 antitoxin, Leu^{59}, Trp^{60}, Asp^{61}, and Lys^{64} (LWDK). To reduce the complexity of our library, we only allowed residues at each library position that are commonly found in naturally occurring ParD homologs (see Experimental Procedures). The resulting library has a theoretical diversity of 9,360 variants, with 12, 6, 13, and 10 possible residues encoded at the four respective positions of the library (Figure 4A). Deep-sequencing of the relevant region in *parD3* in the initial library revealed that >98% of the predicted variants were represented by at least 10 reads and >94% had at least 100 reads (Figure S4A). Measurements of read numbers were highly reproducible between replicates ($R^2 > 0.99$, Figure S4B).

To assess the ability of each ParD3 variant to bind and antagonize ParE3, we co-transformed *E. coli* with the ParD3 library and an inducible ParE3 vector. When cultured in conditions that do not induce ParD3, cell growth arrested within 200 min after inducing the ParE3 toxin (Figure 4B). In contrast, when the ParD3 library was expressed, growth slowed after inducing the toxin but eventually resumed, suggesting that some fraction of the population could neutralize ParE3 toxicity (Figure 4B). To determine which mutants neutralized ParE3 and hence were enriched during the course of this experiment, we harvested samples every 100 min and deep-sequenced the relevant region of *parD3*. We observed large changes in the frequency of individual variants over this time course (Figure S4C). For example, the variant containing the wild-type ParD3 residues (LWDK) was enriched ~6-fold, whereas variants with frameshift mutations in *parD3,* which are presumably non-functional, were depleted ~7-fold (Figure S4C). To validate the functionality of variants inferred from this competitive growth assay, we isolated six mutants that exhibited different frequency dynamics following toxin induction (Figure 4C). We tested these six mutants individually using our toxicity-rescue assay and found clear agreement between the change in the frequency of each variant in the library and its individual plating efficiency (Figure 4D).

To quantify differences in variant behavior during competitive growth, we generated a linear fit to the frequencies of each mutant as a function of time, and then calculated the log-fold expansion of each mutant relative to the rest of the population, producing a raw fitness value (W_{raw}) for each mutant. We then transformed these raw fitness values such that the *W* value for frameshift variants was 0 and the *W* value for the wild-type (LWDK) sequence was 1; the resulting distribution of *W* values ranged from −0.04 to 1.13 and was highly reproducible between biological replicates (Figure 4E, $R^2 = 0.98$). We found a total of 252 variants with W values > 0.5, representing 2.7% of the total (Figure 4F). This set included the wild-type combination of residues (LWDK) and 31 single, 189 double, and 31 triple mutants relative to the wild-type sequence (Figure S4D). There were no quadruple mutants, as position 60 was invariantly tryptophan. The most common residues in this set as a whole were wild-type. However, the identification of 252 variants that can effectively antagonize ParE3 indicates a substantial degree of functional degeneracy in the ParD3 interfacial residues.

Next, to assess the ability of each ParD3 variant to antagonize the non-cognate toxin ParE2, we repeated the competitive growth experiment but co-transformed *E. coli* with our ParD3 library and an inducible ParE2 vector. As before, we observed growth rescue following ParD3 library expression with large changes in the frequency of individual variants over time (Figures 4G and S4E). However, the frequency changes observed here differed from those observed against the cognate toxin ParE3. For example, a variant containing the specificity residues found in the native ParD2 antitoxin, AWIL, was enriched in the ParD3 library screened against ParE2 but was depleted when screened against ParE3 (Figures 4C and 4H). We quantified variant fitness as before and found a total of 151 variants (1.6% of the total) capable of antagonizing ParE2 with W values > 0.5 (Figures 4I and S4E). The most common residues were Ala^{59}, Trp^{60} (invariant), Leu^{61}, and Leu^{64}. However, we noted important differences between variants reactive against ParE2 and ParE3, particularly at the last two variable positions in our library. ParE2-specific variants tended to have small hydrophobic or positively charged residues at position 61, whereas ParE3-specific variants favored negatively charged residues at this position (Figures 4F and 4I). Additionally, ParE2-specific variants were more likely to contain small hydrophobic residues at position 64, whereas ParE3-specific variants tended to have positively charged residues (Figures 4F and 4I).

Mutational Paths That Reprogram Specificity Tend to Involve Promiscuous Variants

To more systematically probe the sequence space governing the specificity of ParD3, we generated a scatterplot of ParD3 variant fitness when screened against the ParE2 or ParE3 toxin (Figure 5A). This analysis revealed variants spanning all ranges of fitness, including those capable of antagonizing ParE2, ParE3, or both toxins simultaneously. We identified a total of 31 promiscuous variants ($W > 0.5$ for both toxins), which represents a subset of the 252 ParE3-reactive and 151 ParE2-reactive variants (Figure 5B). We then grouped variants by specificity class (Figure S5A) and found that the promiscuous variants, such as LWEL, tended to harbor sequence elements from both ParD3 and ParD2, often with negatively charged residues at position 61 (ParD3-like) and aliphatic residues at position 64 (ParD2-like) (Figure 5C).

To visualize the connectivity of functional variants in sequence space, we created a force-directed graph where individual nodes represent functional variants with lines connecting variants that differ by a single amino acid (Figure 5D). Node sizes increase with greater connectivity and node colors represent the specificity class of a given variant (Figure 5D). The resulting graph was densely interconnected but generally grouped variants based on their specificity. The average number of edges per node, or degree, was 17.8 and ranged from 7 to 31. However, we noted that the average number of edges per node was 23% higher for promiscuous variants than for variants specific for ParE2 or ParE3 (Figure 5E). We also generated a force-directed graph in which edges represent variants that differ by a single-nucleotide substitution, following the standard genetic

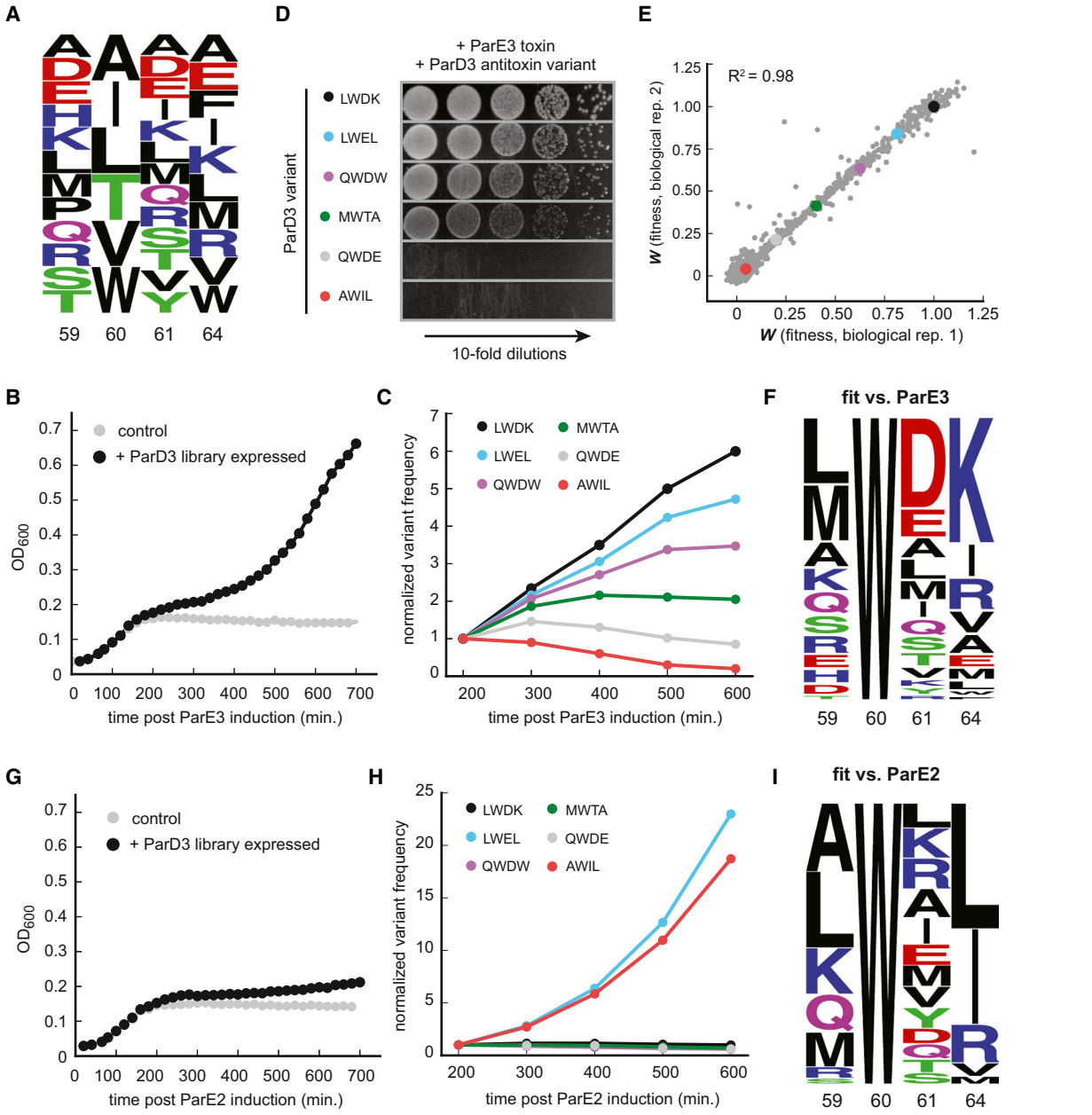


Figure 4. High-Throughput Mapping of Mutant Fitness at Co-evolving Interface

(A) Composition of the ParD3 antitoxin library at the four variable positions.

(B) Library growth following ParE3 toxin induction.

(C) Frequency changes over time for the indicated ParD3 variants following ParE3 induction.

(D) Testing of individual variants from (C) using the toxicity rescue assay. 10-fold serial dilutions were plated from cultures expressing the ParD3 variant indicated and the ParE3 toxin.

(E) Two biological replicates of fitness measurements derived from screening the ParD3 library against the ParE3 toxin.

(F) Frequency logo for ParD3 library variants with high fitness against ParE3 ($W_{E3} > 0.5$).

(G) Library growth following induction of the non-cognate ParE2 toxin.

(H) Frequency changes over time for the indicated ParD3 library variants.

(I) Frequency logo for ParD3 library variants with high fitness against ParE2 ($W_{E2} > 0.5$).

Also see Figure S4.

code (Figure S5B). For this graph, promiscuous variants were, on average, 31% more connected to other nodes than their ParE2- or ParE3-specific counterparts (Figure 5E). This increased connectivity of promiscuous variants was highly significant for both amino acid and nucleotide graphs, as it was lost when the edges of each graph were randomly shuffled ($p < 10^{-4}$,

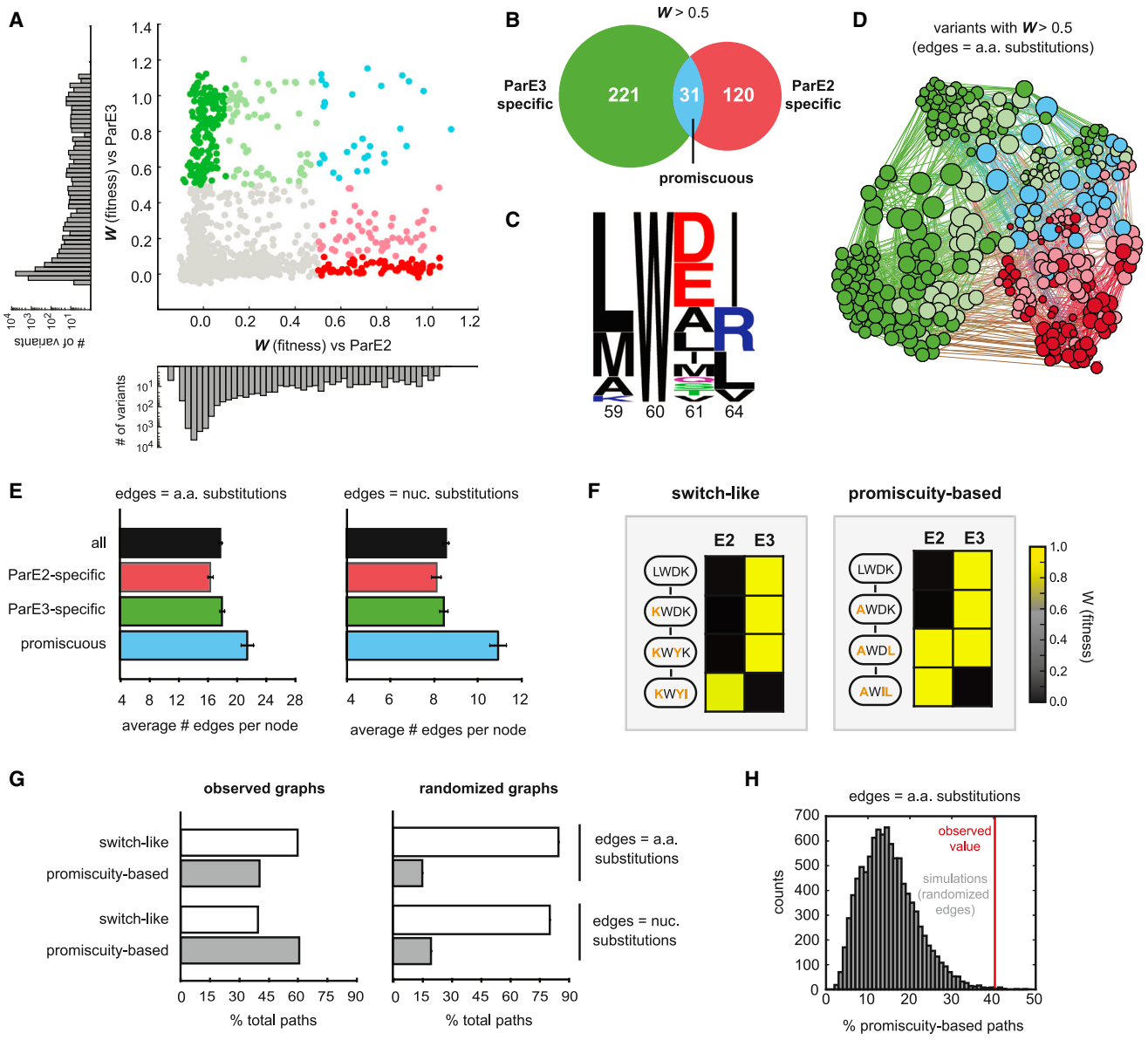


Figure 5. Specificity-Reprogramming Paths Are Highly Enriched for Promiscuous Variants

(A) Fitness of ParD3 variants against ParE2 and ParE3. Green, specific for ParE3; blue, capable of antagonizing both ParE2 and ParE3; red, specific for ParE2. Histograms of fitness values against ParE2 and ParE3 are shown.

(B) Venn diagram of ParD3 variants reactive against ParE3, ParE2, or both.

(C) Frequency logo of promiscuous ParD3 variants ($W_{E2} > 0.5$, $W_{E3} > 0.5$).

(D) Force-directed graph of all ParD3 variants reactive against ParE3 or ParE2 ($W > 0.5$). Nodes represent individual variants and edges represent single amino-acid substitutions. Node size scales with increasing degree and color corresponds to the specificity classes in (A).

(E) Average number of edges per node for the indicated categories of ParD3 variants. Error bars indicate SEM.

(F) Examples of "switch-like" and "promiscuity-based" mutational paths from an E3-specific variant to an E2-specific variant with the fitness against each variant color-coded based on the scale shown.

(G) Left, percentage of "switch-like" and "promiscuity-based" paths from the wild-type ParD3 sequence (LWDK) to each of the 66 ParE2-specific variants ($W_{E2} > 0.5$, $W_{E3} < 0.1$). Right, same as left panel but for 10,000 simulations in which the graph edges were randomly shuffled while keeping the total edge count and degree distribution constant. Error bars represent SEM.

(H) Histogram representing percentage of "promiscuity-based" paths in 10,000 edge shuffling simulations; red line indicates percentage for the observed amino acid graph.

Also, see Figure S5.

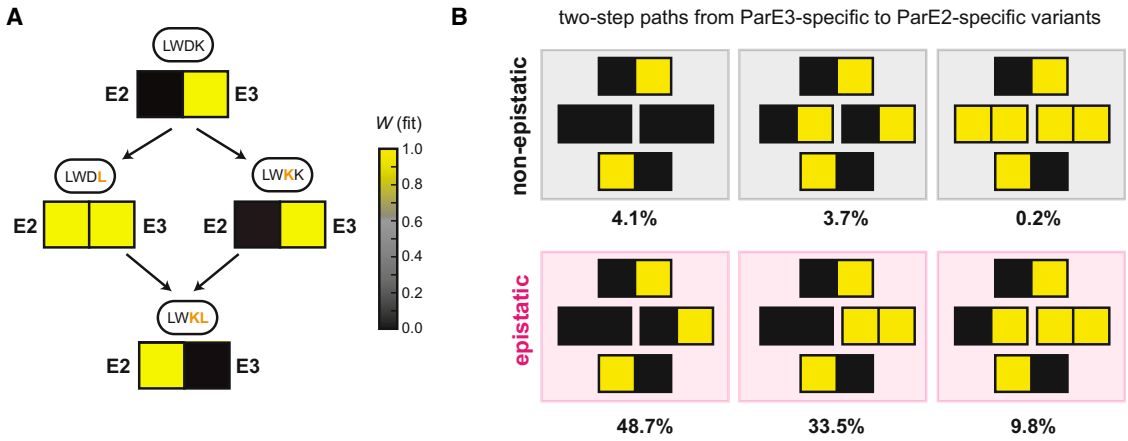


Figure 6. Mutational Order Dictates Specificity Class of Intermediate Variants

(A) Mutational paths from LWDK to LWKL for ParD3 with fitness of each variant against ParE2 and ParE3 shown as a heatmap: yellow, high fitness; black low fitness.

(B) The six path types that reprogram ParD3 specificity in two mutational steps. Percentage of mutational paths in each category is indicated for a threshold of 0.5 used to define a positive interaction.

Also see Figure S6.

Figures S5C and S5D). The high connectivity of promiscuous variants was even more pronounced with a more stringent definition of specificity (Figure S5E).

The dense connectivity of promiscuous variants suggested that mutational paths that change ParD3 specificity (from ParE3-specific to ParE2-specific, or vice versa) tend to travel through promiscuous intermediates. To test this hypothesis, we first defined two types of specificity-reprogramming paths. Note that for the following analysis, we exclude paths in which ParD3 fails to interact with both ParE3 and ParE2 (also see Discussion). The first class of paths are "switch-like" and only involve intermediates that are specific for ParE2 or ParE3, whereas the second class of paths are "promiscuity-based" and travel through at least one intermediate that can inhibit both ParE2 and ParE3 (Figure 5F). To determine whether paths that change the interaction specificity of ParD3 tend to be switch-like or promiscuity-based, we identified all shortest mutational paths from the wild-type ParD3 variant (LWDK) to each of the 66 variants that are highly specific for ParE2 ($W_{E2} > 0.5$, $W_{E3} < 0.1$; Figure S5A); for this analysis, each mutational step involved a single amino-acid substitution. We found a total of 370 shortest paths, of which 40% involved a promiscuous intermediate (Figure 5G). The percentage of paths via promiscuous intermediates increased to 61% when considering only paths that involve single-nucleotide substitutions (Figure 5G).

To determine whether the number of paths that involve promiscuous variants is greater than would be expected by chance, we generated graphs in which the edges were randomly shuffled, and again calculated the percentage of each class of paths from ParD3 (LWDK) to the ParE2 highly specific variants. For these graphs with randomized edges, the percentage of paths involving promiscuous intermediates dropped to 15% for the amino acid neighbor graph and 20% for the nucleotide neighbor graph (Figures 5G and 5H). Thus, the enrichment of promiscuity-based paths in the observed graphs is significant ($p < 0.005$) (Figures 5G, 5H, and S5F). Collectively, our results demonstrate the dense connectivity of functional variants in the sequence space governing ParD-ParE interaction specificity and reveal that specificity-reprogramming paths are highly enriched for those that involve promiscuous variants, which may facilitate the evolution of ParD-ParE systems with new specificities.

Epistasis: Mutational Order Dictates Specificity Class of Intermediate Variants

Inspection of the paths connecting ParD3 variants with different specificities indicated that the third and fourth library positions, residues 61 and 64 in ParD3, contribute significantly to the insulation of the ParD-ParE system. For instance, the wild-type residue combination in ParD3, LWDK, renders it specific for binding to ParE3, whereas the double-mutant variant LWKL is specific for ParE2. Strikingly, however, the two possible paths connecting LWDK and LWKL are in different classes (Figure 6A). A single ParD3 substitution (K64L in LWDL) resulted in promiscuous binding to ParE2 and ParE3, whereas a second substitution in this background (D61K in LWKL) resulted in specificity for ParE2 (Figure 6A). In contrast, incorporating these substitutions in the reverse order, D61K and then K64L, resulted in a switch-like change in specificity in which the initial D61K substitution retained specificity for ParE3, but then enabled the subsequent K64L substitution to produce a ParE2-specific antitoxin (Figure 6A). These results underscore how a small number of mutations can fully reprogram protein-protein interaction specificity and demonstrate that the order of mutations can strongly affect whether the path to a new specificity state involves a promiscuous intermediate or a rapid switch.

Our finding that changes in specificity can depend strongly on the order of substitutions represents a form of epistasis, broadly defined as cases where the functional effect of individual substitutions is context-dependent rather than additive and independent (Lehner, 2011). To more broadly quantify this epistasis for

the ParD3 interfacial residues, we first defined six types of specificity-reprogramming paths that involve two amino-acid substitutions (Figure 6B). Three of the six path types are epistatic with the two intermediates having different specificities, implying that substitution order influences changes from ParE3 to ParE2 specificity. We quantified the path type for each case in which two substitutions reprogram ParD3 from being specific for ParE3 ($W_{E3} > 0.5$, $W_{E2} < 0.5$) to being specific for ParE2 ($W_{E3} < 0.5$, $W_{E2} > 0.5$) and found a total of 2,653 such cases, of which 92% were epistatic (Figure 6B). The percentage of epistatic paths was robust to the threshold used for defining positive interactions (Figures S6A and S6B). Taken together, our results highlight the pervasive effects of epistasis on ParD function. Although studies of epistasis typically consider the interdependence of individual substitutions with respect to protein folding or a single-protein function (Kondrashov and Kondrashov, 2015; Lehner, 2011), our findings indicate that epistasis can also manifest at the level of interaction specificity. This form of epistasis may significantly impact the evolution of new ParD-ParE systems. Promiscuous intermediates enable a change in protein-protein interaction specificity without passing through a non-functional state, in which a liberated toxin would suppress growth and proliferation (Figure 1A). Thus, the epistasis documented here may fundamentally restrict mutational trajectories during evolution to those involving promiscuous intermediates.

Mutational Trajectories to an Orthogonal ParD3-ParE3 Pair

Thus far, we have considered changes to one side of the ParD-ParE interface. To probe how the interaction specificity of a ParD-ParE protein pair coevolves, we sought to generate a variant of the toxin ParE3 that does not interact with ParD3, and then select ParD3 variants from our library that can neutralize this novel toxin. To this end, we generated a variant of the toxin, called ParE3*, that retains toxicity but is incapable of binding to the ParD3 antitoxin. In particular, we mutated five ParE3 positions (Arg^{54}, Arg^{58}, Ala^{61}, Met^{63}, and Leu^{72}, or RRAML) that strongly covary with the specificity residues in ParD3. We mutated RRAML→VEIRF, as each individual variant residue was frequently observed in ParE3 homologs and was chemically different from the corresponding wild-type residue (Figure S7A). As expected, we found that ParE3* retained toxicity but was no longer neutralized by ParD3 (Figure 7A).

To determine whether variants in the ParD3 library neutralized ParE3*, we performed a competitive growth experiment following co-transformation. As before, we converted changes in variant frequencies to fitness values, which were highly reproducible ($R^2 = 0.96$, Figure S7B). Sequence analysis of the high-fitness mutants ($W > 0.5$) revealed large differences in amino-acid preferences at positions 60 and 61 relative to those shown above (Figures 4F and 7B). In particular, for the ParD3 variants that neutralized ParE3*, the invariant Trp^{60} was replaced by Ile/Val/Leu and the strong preference for a negatively charged residue at position 61 was replaced by positively charged or neutral residues (Figures 4F and 7B). One of the high-fitness variants with specificity residues LIAK, renamed ParD3*, no longer neutralized ParE3 but robustly interacted with ParE3* (Figure 7C). Taken together, our results indicate that mutations in the specificity residues of ParD3 and ParE3 are sufficient to create an orthogonal, interacting protein pair.

Our results indicate that mutational paths leading to a change in ParD specificity tend to pass through promiscuous intermediates (Figure 5). Thus, we wanted to determine whether mutational paths between the wild-type ParD3-ParE3 and the orthogonal ParD3*-ParE3* systems also pass through promiscuous intermediates, thereby changing the specificity of both proteins without disrupting their interaction. We therefore generated variants of ParE3 containing all possible subsets of the substitutions in ParE3* (32 mutants) and variants of ParD3 containing all possible subsets of the substitutions in ParD3* (4 mutants). We then co-transformed each possible pairing of ParD3 and ParE3 variants (128 pairs total) into *E. coli* and assessed interaction using the toxicity-rescue assay (Figure 7D). Interestingly, 90 of the 128 pairs of ParD3 and ParE3 variants were capable of interacting, likely because most (17 of 32) of the ParE3 variants were promiscuous, which we define as interacting strongly with both ParD3 and ParD3* (Figure 7D).

To determine whether paths between the wild-type and insulated ParD-ParE pairs tend to pass through promiscuous intermediates, we first enumerated the total number of trajectories between these systems. Assuming one residue is changed per step and no reversions are considered, there are 5,040 paths from ParD3-ParE3 to the orthogonal ParD3*-ParE3* pair; of these paths, 1,030 retain functionality at each intermediate step. Strikingly, we found that all of these 1,030 functional paths passed through at least one promiscuous intermediate of ParE3 with an average of five promiscuous ParE3 intermediates per path (Figure S7C). The prevalence of these promiscuous states may enable the ParD-ParE system to readily evolve a new interaction specificity. An initial broadening of ParE3 specificity enables the movement of ParD3 in sequence space, followed by a narrowing of ParE3 specificity in the final step (Figure 7E). By contrast, mutational paths in which a substitution in either ParD or ParE yields a "switch-like" change in specificity would, by definition, be broken until a second substitution restores the interaction. Thus, our results support the notion that the coevolution and expansion of the ParD-ParE family occurs through promiscuous intermediates.

DISCUSSION

Mutational Trajectories and the Coevolution of Protein-Protein Interactions

Interacting proteins coevolve, and the identification of coevolving amino acids in two proteins can often help to pinpoint the residues that mediate their interaction. Such analyses are typically predicated on the idea that a mutation in one protein that disrupts an interaction then drives selection of a compensatory mutation in the partner, thereby restoring the interaction (Figure 1A). However, this model implies that organisms tolerate (at least transiently) a non-functional, or less functional, interaction, which seems unlikely if the protein-protein interaction is essential for viability. Our results provide a solution to this conundrum, demonstrating experimentally how interacting proteins can coevolve and acquire new specificity by having one of the proteins pass through a promiscuous intermediate (Figure 1B). For

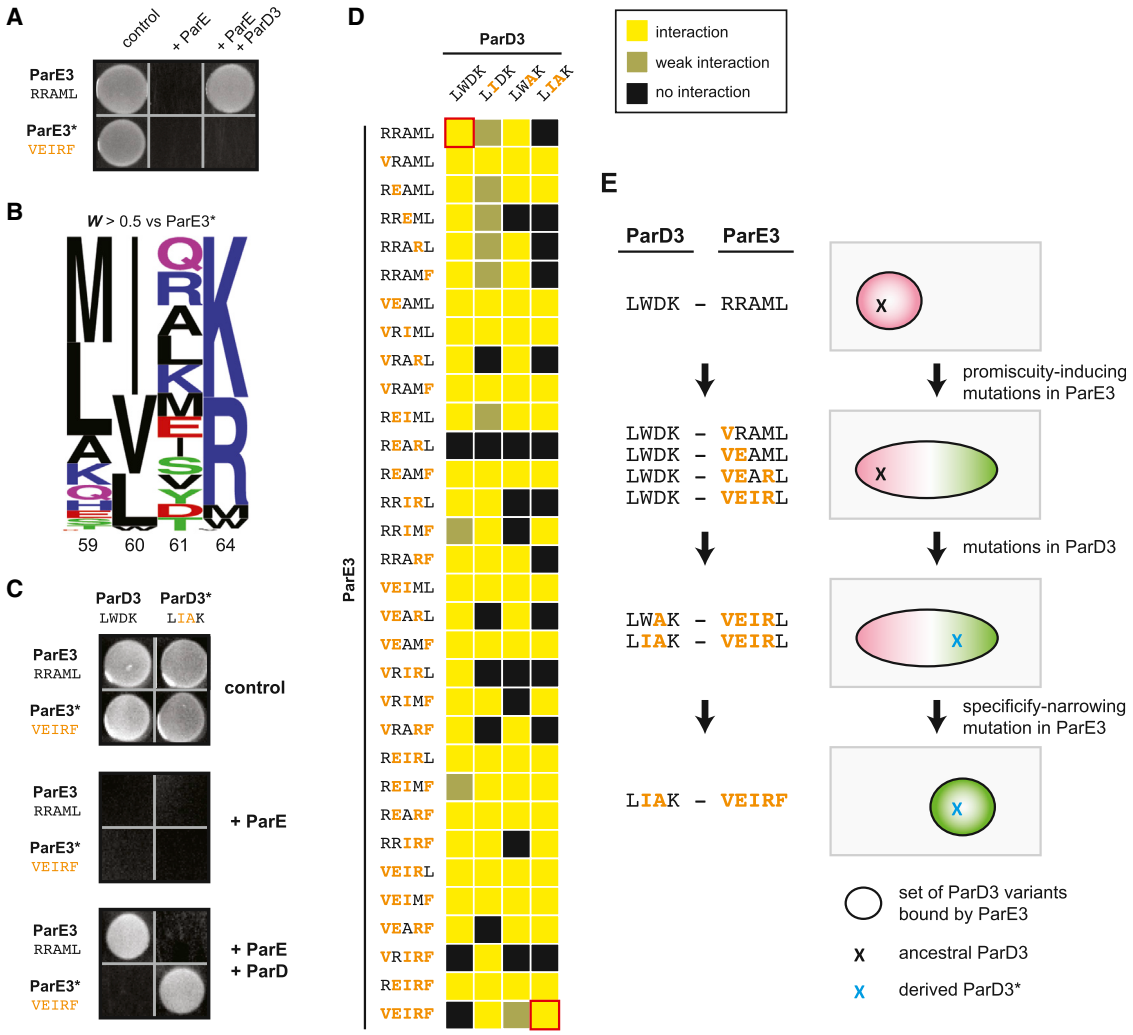


Figure 7. Mutational Trajectories to an Orthogonal ParD3*-ParE3* Pair

(A) ParE3* is insulated from antitoxin ParD3. A plasmid containing either ParE3 or ParE3* was co-transformed into *E. coli* with a plasmid expressing ParD3, and cells were plated on medium that induces or represses expression of the toxin and antitoxin.

(B) Frequency logo for ParD3 library variants with high fitness against ParE3* ($W_{E3}^{*} > 0.5$).

(C) ParE3*-ParD3* is insulated from the wild-type ParD3-ParE3 pair.

(D) Toxicity-rescue interaction assays for all ParD3 and ParE3 mutant combinations. Top left, wild-type ParD3-ParE3 pair; bottom right, orthogonal ParD3*-ParE3* pair. Promiscuous ParE3 intermediates are those capable of interacting with both ParD3 and ParD3*.

(E) Example of a series of single substitutions that lead to the insulated ParE3*-ParD3* system while retaining the toxin-antitoxin interaction at each step by first expanding the specificity of ParE3, followed by changes in ParD3, and finally by restricting the specificity of ParE3.

Also see Figure S7.

instance, a mutation in an antitoxin can initially broaden its specificity; the toxin can then accumulate a mutation that moves it in sequence space but retains its interaction with the antitoxin. A subsequent substitution in the antitoxin can then narrow its specificity to include the mutated toxin and exclude the original form. The net result is a change in specificity without disruption of the protein-protein interaction, which is critical as a disruption at any step would liberate a toxin that prevents growth and proliferation. This model for protein coevolution involves a minimum of three instead of two mutations but means that the protein-protein interaction is functional at each step. Thus, such mutational trajectories could be entirely neutral but importantly would retain a pairwise-coevolution signature in multiple sequence alignments.

Our systematic identification of ParD3 variants that can antagonize ParE3, ParE2, or both revealed an abundance of promiscuous variants in sequence space that are, on average, more highly connected to other functional variants than are specific variants. Consequently, the mutational trajectories that reprogram the specificity of ParD3 frequently involve promiscuous intermediates (Figures 5F and 5G). The high frequency of mutational paths involving promiscuous intermediates was seen when

considering transitions in ParD3 from being specific for ParE3 to specific for ParE2, and even more so when considering mutations on both sides of the interface. We assessed the complete set of mutational trajectories between the wild-type ParD3-ParE3 and the orthogonal ParD3*-ParE3* by testing 128 pairwise interactions between all possible ParD3 and ParE3 mutational intermediates. Strikingly, 17 of the 32 ParE3 intermediate variants were promiscuous, or capable of interacting with both the ParD3 and ParD3* variants (Figure 7). Consequently, all of the functional paths between ParD3-ParE3 and ParD3*-ParE3* involved at least one promiscuous intermediate, with most involving more than five (Figure 7). Our results thus suggest that promiscuous variants of ParD and ParE are abundant in sequence space and that promiscuity-enabling mutations can facilitate the evolution of new interaction specificities while still using the same set of interfacial residues.

A similar principle may apply to other protein-protein interactions throughout biology, even those not involving toxic proteins. The disruption of a given protein-protein interaction could prevent the execution of an essential cellular function or lead to an unwanted, detrimental interaction with another protein, thus favoring coevolutionary trajectories that retain function at each step. This same principle may also underlie the coevolution of transcription factors and their DNA binding sites. The evolutionary history of a steroid hormone receptor and its recognition element was recently reconstructed including the analysis of a possible ancestral state of the steroid receptor and mutational intermediates separating it from extant states (Anderson et al., 2015). Several of the intermediates were promiscuous and may have facilitated coevolution of the receptor and its recognition element toward a new specificity without disrupting the interaction. However, that study only considered mutational intermediates containing residues present in the ancestral or derived states, and our analyses of the ParD-ParE interface suggest that promiscuous intermediates can also involve substitutions that appear in neither the ancestral nor the derived states.

Like many protein families, toxin-antitoxin systems can expand through duplication and divergence. The duplication of a toxin-antitoxin system could allow one of the protein pairs to wander unconstrained in sequence space toward a new interaction specificity via switch-like paths that involve non-functional intermediates. After a duplication, one antitoxin could accumulate interaction-disrupting substitutions while its toxin is still inhibited by the other antitoxin. The toxin could then subsequently mutate to restore an interaction with the derived antitoxin. However, this scenario assumes that the evolving antitoxin does not, in the intermediate state, interact inappropriately with other proteins, and it assumes that the other antitoxin is produced at sufficiently high levels to inhibit 2-fold more toxin, i.e., that there is normally a significant excess of free antitoxin, which may not be the case. Determining whether and when switch-like or promiscuous paths are followed will require careful reconstructions of toxin-antitoxin evolution.

High-Throughput Mapping of Protein Interaction Specificity

Deep mutational scanning via next-generation sequencing is a relatively new approach for interrogating the relationship between protein sequence and function, including folding, enzymatic activity, or the binding of a target protein or RNA (Fowler and Fields, 2014). These studies have begun to reveal the functional degeneracy of proteins by examining all, or nearly all, possible single mutants of a given protein. Similar approaches have also been used to probe subsets of all possible double and higher-order mutants (Melamed et al., 2013) or to systematically probe all possible mutants at a limited set of positions (Podgornaia and Laub, 2015).

Deep mutational scans have been focused primarily on how mutations alter a single function or protein interaction. One study examined the ability of a PDZ domain to interact with both a cognate and non-canonical peptide ligand (McLaughlin et al., 2012), but only queried single-point mutants. However, the interaction specificity of a protein is a distributed property of multiple amino acids, and the prevalence of epistasis means that the behavior of multiple mutations is difficult to infer from the properties of the corresponding single mutants. We queried a diverse library of ParD3 variants harboring multiple mutations of key specificity residues against two separate proteins: the cognate toxin ParE3 and the non-cognate toxin ParE2. This focused library approach was possible as the specificity of ParD is largely determined by a small number of interfacial residues (Figure 3). Our approach yielded a high-density map of the sequence space of ParD3 that underpins its substrate interaction specificity (Figures 5A–5D). From these data, we uncovered the residues in ParD3 most responsible for its selective binding of one toxin over another (Figures 4F and 4I). We found that three positions (60, 61, and 64) primarily dictate specificity, with substitutions at two sites (61 and 64) sufficient to switch ParD3 from antagonizing ParE3 to ParE2, and substitutions at an overlapping set of sites (60 and 61) sufficient to switch ParD3 from antagonizing ParE3 to ParE3*. As noted, our results also demonstrated the existence of many residue combinations that promote a promiscuous state of ParD3 or ParE3. Mutations that render proteins more promiscuous, with respect to catalytic activities or binding partners, has been noted anecdotally (Aharoni et al., 2005; Bloom and Arnold, 2009), but the prevalence of such states and, importantly, their accessibility from more specific, wild-type states has never been mapped in a comprehensive manner.

By building and screening libraries harboring multiple mutations, our work also sheds new light on protein epistasis and the non-additive relationship of individual substitutions. Epistasis has been well documented but is typically assessed with respect to a single-protein function. By contrast, the epistasis documented here for ParD3 pertains to its specificity and interaction with two different proteins, revealing interdependencies that would be missed when considering only a single function. For instance, consider the example in Figure 6A where ParD3 transitions from the E3-specific residues LWDK to the E2-specific residues LWKL. With respect to antagonizing the toxin ParE3, the two single mutants, LWDL and LWKK, are each functional. However, with respect to toxin ParE2, LWDL is functional whereas LWKK is not, reflecting a non-additive relationship between the two substitutions leading to the double mutant LWKL. This type of epistasis may, like other forms of epistasis, restrict the evolution of ParD-ParE systems, which likely follows

mutational paths that involve promiscuous states, as discussed above.

Interaction Specificity of Toxin-Antitoxin Systems

The specificity of interactions in bacterial toxin-antitoxin systems had previously been unclear, with some reports indicating that these protein-protein interactions are specific (Fiebig et al., 2010) and others suggesting that TA systems form large, cross-reactive networks (Yang et al., 2010; Zhu et al., 2010). Here, by performing a systematic assessment of interaction specificity for a TA family, we found that ParD antitoxins typically exhibit an exquisite preference for binding to their co-transcribed ParE toxins, forming exclusive, cognate pairs. Of 180 non-cognate pairings tested, we found cross-talk in only 11 cases (Figure 2) and, importantly, no cross-talk was observed for non-cognate pairs present in the same species.

The high degree of protein-protein interaction specificity observed for the ParD-ParE family is similar to that observed for other large, paralogous protein families (Newman and Keating, 2003; Skerker et al., 2008; Stiffler et al., 2007; Zarrinpar et al., 2003). The specificity of many of these paralogous families has been attributed to selection against detrimental cross-talk (Capra et al., 2012; Zarrinpar et al., 2003), raising the possibility that the ParD-ParE family may be under similar selective pressures. However, the biological rationale for maintaining the specificity of TA systems is unclear, and will require a deeper understanding of the function of these systems in bacterial physiology.

Final Perspective

In sum, our work provides a rationale and molecular basis for how protein interaction specificity can change and how two proteins can coevolve without involving non-functional intermediates. Mutations that produce promiscuity have been described for a variety of proteins, but the frequency of such mutations and their accessibility from more specific states had been unclear. Our results indicate that, at least for ParD3 and likely other proteins, promiscuous mutants are prevalent and easily reached from the wild-type sequence through a single mutation. The prevalence of promiscuous intermediates may facilitate the expansion of toxin-antitoxin systems and, more broadly, other paralogous protein families.

EXPERIMENTAL PROCEDURES

ParD3-ParE Structure Analysis

For details on the structural analysis of *M. opportunistum* ParD3 and ParE3, see Supplemental Experimental Procedures.

Identification of Coevolving Residues

Coevolving residues in the ParDE family were identified using GREMLIN at http://gremlin.bakerlab.org. Input sequences were ParD3 and ParE3 from *M. opportunistum*, and we set the number of iterations to four and the E-value cutoff to 1E-04. To identify specificity residues, we isolated all residue pairings that had a scaled coupling score greater than 1.25. To identify supporting residues, we performed the following iterative procedure using a score cutoff of 1.25: (1) identify residues within ParD or ParE that covary with the specificity residues; (2) identify residues within ParD or ParE that covary with either the specificity residues or the supporting residues identified in step (1); (3) repeat step (2) until no new supporting residues are identified.

ParD3 Library Construction and Analysis

For details on construction of the ParD3 library, see the Supplemental Experimental Procedures. To assess the ability of each ParD3 variant to antagonize different ParE toxins, *E. coli* cells harboring the ParD3 plasmid library were electroporated with a plasmid containing an arabinose-inducible copy of the ParE toxin. Cells were grown out overnight in 200 ml M9L supplemented with 0.4% glucose and antibiotics. The following day, cells were spun down, washed in 50 ml of M9L, and re-suspended at an OD of 0.03 in 500 ml of M9L supplemented with 100 μM IPTG (to induce the ParD3 library) and antibiotics. Cells were grown out at 37°C with shaking for 100 min, and then ParE toxin expression was induced by the addition of 0.2% arabinose. Cell density was measured every 20 min and samples (50 ml) were taken every 100 min, pelleted, and frozen at −20°C. Competitive liquid growth assays were performed in duplicate. Plasmid DNA was extracted and used as template for PCR (20 cycles) with custom barcoded primers containing Illumina flowcell adaptor sequences. Samples were sequenced on an Illumina HiSeq and then filtered, counted, and converted to fitness values as described in the Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.055.

AUTHOR CONTRIBUTIONS

Crystallization experiments performed by J.H. and S.C. Protein chimeras in Figure 3 generated by T.N.P. Toxicity-rescue assays in Figure 7 performed by B.S.P. All other experiments performed by C.D.A. C.D.A. and M.T.L. designed experiments, analyzed data, and wrote the paper.

ACKNOWLEDGMENTS

We thank R. Sauer, A. Murray, and the Laub laboratory for discussions and comments on the manuscript. We acknowledge S. Ovchinnikov and C. Bahl for valuable discussions on GREMLIN. This work supported by a NIH grant (5R01GM082899) to M.T.L. who is also an Investigator of the Howard Hughes Medical Institute.

Received: July 23, 2015
Revised: September 11, 2015
Accepted: September 22, 2015
Published: October 15, 2015

REFERENCES

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., and Tawfik, D.S. (2005). The 'evolvability' of promiscuous protein functions. Nat. Genet. *37*, 73–76.

Anderson, D.W., McKeown, A.N., and Thornton, J.W. (2015). Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. eLife *4*, e07864.

Bloom, J.D., and Arnold, F.H. (2009). In the light of directed evolution: pathways of adaptive protein evolution. Proc. Natl. Acad. Sci. USA *106* (*Suppl 1*), 9995–10000.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222–232.

Dalton, K.M., and Crosson, S. (2010). A conserved mode of protein recognition and binding in a ParD-ParE toxin-antitoxin complex. Biochemistry *49*, 2205–2215.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. *6*, 678–687.

Fiebig, A., Castro Rojas, C.M., Siegal-Gaskins, D., and Crosson, S. (2010). Interaction specificity, toxicity and regulation of a paralogous set of ParE/RelE-family toxin-antitoxin systems. Mol. Microbiol. *77*, 236–251.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Hallez, R., Geeraerts, D., Sterckx, Y., Mine, N., Loris, R., and Van Melderen, L. (2010). New toxins homologous to ParE belonging to three-component toxin-antitoxin systems in Escherichia coli O157:H7. Mol. Microbiol. *76*, 719–732.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. USA *110*, 15674–15679.

Kondrashov, D.A., and Kondrashov, F.A. (2015). Topological features of rugged fitness landscapes in sequence space. Trends Genet. *31*, 24–33.

Kuriyan, J., and Eisenberg, D. (2007). The origin of protein interactions and allostery in colocalization. Nature *450*, 983–990.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends Genet. *27*, 323–331.

Leplae, R., Geeraerts, D., Hallez, R., Guglielmini, J., Drèze, P., and Van Melderen, L. (2011). Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. Nucleic Acids Res. *39*, 5513–5525.

McLaughlin, R.N., Jr., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA *19*, 1537–1551.

Newman, J.R., and Keating, A.E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science *300*, 2097–2101.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife *3*, e02030.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673–677.

Ramage, H.R., Connolly, L.E., and Cox, J.S. (2009). Comprehensive functional analysis of Mycobacterium tuberculosis toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution. PLoS Genet. *5*, e1000767.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043–1054.

Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364–369.

Yamaguchi, Y., Park, J.H., and Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. Annu. Rev. Genet. *45*, 61–79.

Yang, M., Gao, C., Wang, Y., Zhang, H., and He, Z.G. (2010). Characterization of the interaction and cross-regulation of three Mycobacterium tuberculosis RelBE modules. PLoS ONE *5*, e10672.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676–680.

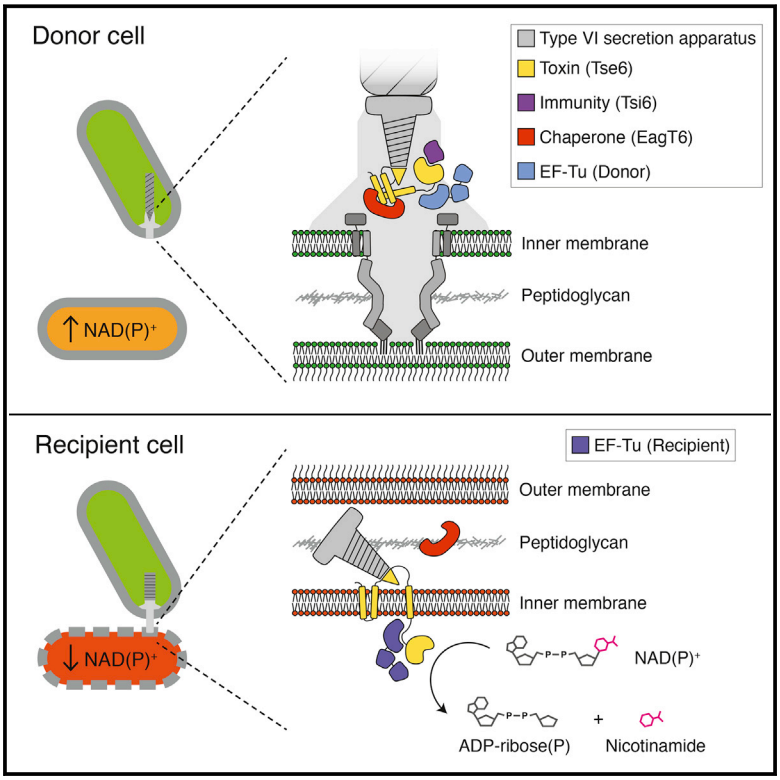
Zhu, L., Sharp, J.D., Kobayashi, H., Woychik, N.A., and Inouye, M. (2010). Noncognate Mycobacterium tuberculosis toxin-antitoxins can physically and functionally interact. J. Biol. Chem. *285*, 39732–39738.

Article

Cell

An Interbacterial NAD(P)$^+$ Glycohydrolase Toxin Requires Elongation Factor Tu for Delivery to Target Cells

Graphical Abstract



Authors

John C. Whitney, Dennis Quentin, Shin Sawai, ..., David R. Goodlett, Stefan Raunser, Joseph D. Mougous

Correspondence

mougous@u.washington.edu

In Brief

To protect its niche within a microbial community, *Pseudomonas* secretes a toxin that depletes competing bacterial cells of NAD$^+$ and NADP$^+$. Structures of the toxin and accessory secretory proteins reveal a surprising requirement for the housekeeping protein EF-Tu in toxin delivery and provide mechanistic insight into intercellular protein transport between bacteria.

Highlights

- Type VI secretion effector Tse6 acts by depleting bacteria of NAD$^+$ and NADP$^+$
- Entry of Tse6 into target cells requires its binding to elongation factor Tu
- Tse6 is a membrane protein that requires a chaperone for intercellular transport
- EM structures reveal the mechanism for Tse6 deployment to recipient cells

Accession Numbers

4ZV0
4ZUY
4ZV4

Whitney et al., 2015, Cell *163*, 607–619
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.027

CellPress

Cell

An Interbacterial NAD(P)$^+$ Glycohydrolase Toxin Requires Elongation Factor Tu for Delivery to Target Cells

John C. Whitney,[1] Dennis Quentin,[2] Shin Sawai,[1] Michele LeRoux,[1] Brittany N. Harding,[1] Hannah E. Ledvina,[1] Bao Q. Tran,[3] Howard Robinson,[4] Young Ah Goo,[3] David R. Goodlett,[3] Stefan Raunser,[2] and Joseph D. Mougous[1,5,*]
[1]Department of Microbiology, University of Washington, Seattle, WA 98195, USA
[2]Department of Structural Biochemistry, Max Planck Institute of Molecular Physiology, Otto-Hahn-Strasse 11, 44227 Dortmund, Germany
[3]Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, MD 21201, USA
[4]Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA
[5]Howard Hughes Medical Institute, Seattle, WA 98195, USA
*Correspondence: mougous@u.washington.edu
http://dx.doi.org/10.1016/j.cell.2015.09.027

SUMMARY

Type VI secretion (T6S) influences the composition of microbial communities by catalyzing the delivery of toxins between adjacent bacterial cells. Here, we demonstrate that a T6S integral membrane toxin from *Pseudomonas aeruginosa*, Tse6, acts on target cells by degrading the universally essential dinucleotides NAD$^+$ and NADP$^+$. Structural analyses of Tse6 show that it resembles mono-ADP-ribosyltransferase proteins, such as diphtheria toxin, with the exception of a unique loop that both excludes proteinaceous ADP-ribose acceptors and contributes to hydrolysis. We find that entry of Tse6 into target cells requires its binding to an essential housekeeping protein, translation elongation factor Tu (EF-Tu). These proteins participate in a larger assembly that additionally directs toxin export and provides chaperone activity. Visualization of this complex by electron microscopy defines the architecture of a toxin-loaded T6S apparatus and provides mechanistic insight into intercellular membrane protein delivery between bacteria.

INTRODUCTION

Bacteria utilize a diverse group of secreted toxins to establish and defend their niche. Among these are the effectors exported by the type VI secretion system (T6SS), which are delivered to target cells in a contact-dependent manner (Hood et al., 2010; LeRoux et al., 2012; Russell et al., 2011). Despite the tremendous number and predicted diversity of T6 effectors, few activities have been ascribed to this important group of proteins.

The majority of characterized T6 effectors act in the periplasm of target Gram-negative cells. Within this compartment, the proteins disrupt essential structures, such as cell-wall peptidoglycan (via amidase and glycoside hydrolase activity), and cellular membranes (via phospholipase and pore-forming activity) (Russell et al., 2014). Although a large number of cytotoxic T6 effectors have been identified, the mechanisms by which they influence recipient cells are not well understood (Fritsch et al., 2013; Hood et al., 2010; Whitney et al., 2014). Indeed, a group of related effectors that exhibit DNase activity are the only such proteins yet characterized (Ma et al., 2014).

Unlike other proteinaceous toxins, such as the colicins, T6S effectors do not possess cell-entry mechanisms. Rather, they transit the T6SS, which breaches the outer membrane of recipient cells and thereby grants its substrates access to the cell interior (Russell et al., 2011). Many components of the T6SS bear structural and functional relatedness to tail proteins of contractile bacteriophage (Silverman et al., 2012). The delivery of T6 effectors into recipient cells has not been directly visualized; however, it is likely that they are propelled into recipient cells during phage-like contraction events of the apparatus (Basler et al., 2012). How T6 effectors are recruited to the secretory apparatus is not completely understood. Evidence suggests at least two genetically distinct mechanisms operate. One subset of T6 effectors requires direct interaction with the interior of ring-shaped phage tail tube-like haemolysin co-regulated proteins (Hcp) for export (Silverman et al., 2013). Hcp proteins themselves are abundantly secreted in a T6-dependent manner, leading to the proposal that these toxins are delivered to recipient cells in complex with Hcp. The relatively low molecular weight of Hcp-associated effectors suggests that interaction with the pore of Hcp places constraints on the size of toxins that can be delivered via this pathway.

A second subset of effectors, including many that are high-molecular-weight, multi-domain proteins, require specific valine-glycine repeat protein G (VgrG) type proteins for export (Hachani et al., 2014; Whitney et al., 2014). VgrG proteins form homotrimeric assemblies that have extensive structural homology with phage tail spike proteins, and, like Hcp, are secreted in a T6-dependent manner. Also analogous to Hcp, the requirement for VgrG proteins in effector export is thought to reflect a physical association of these proteins with cognate effectors. The biochemical basis for VgrG-effector interaction is not well studied; however, modular adaptor domains—present either as domains within the effector protein or as independent polypeptides—appear to mediate binding. One such domain harbors PAAR repeat sequences, which fold into a pyramidal structure that interacts

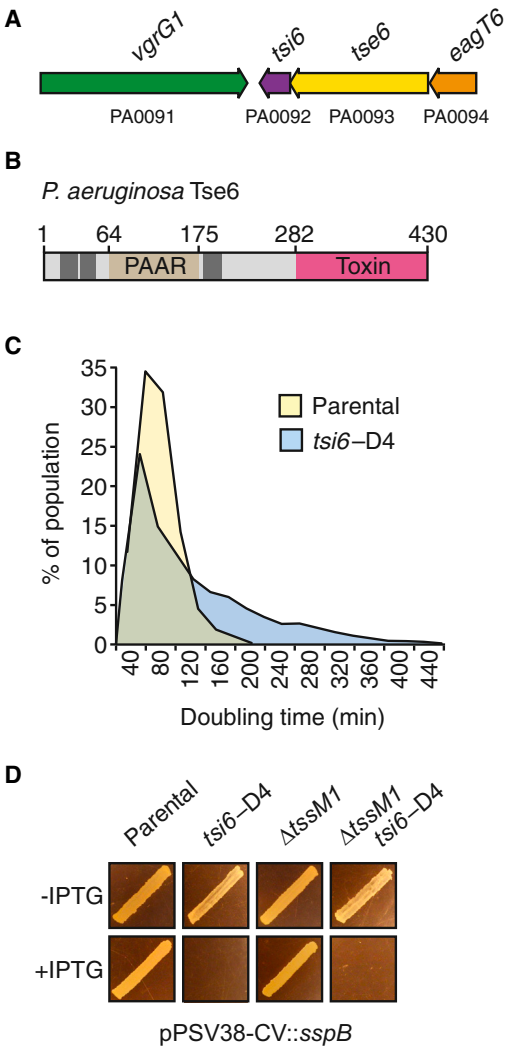


Figure 1. Tse6 Causes Stasis from the Cytoplasm of *P. aeruginosa*

(A) Genomic context of *vgrG1*, *tsi6*, *tse6*, and *eagT6* in *P. aeruginosa* PAO1. Locus tag numbers are provided below each gene. The color of each gene corresponds to the color of its encoded protein shown in subsequent figures.

(B) Domain organization of *P. aeruginosa* Tse6. The boundaries for the PAAR (residues 64–175) and toxin (residues 282–430) domains are indicated. Predicted transmembrane domains are shown as dark gray rectangles.

(C) Intoxication of *P. aeruginosa* by Tse6 severely reduces growth. Data were derived from single-cell analysis of a parental strain (Δ*retS* Δ*sspB* pPSV38::*sspB*) and a derivative depleted of Tsi6 (Δ*retS* Δ*sspB* *tsi6*-D4 pPSV38::*sspB*). Bin size is 20 min and is normalized to total cells (parental, n = 15,042; *tsi6*–D4, n = 5,568).

(D) Tsi6 depletion strains undergo Tse6-based toxicity independent of intercellular toxin delivery by a functional H1-T6SS. Patches of the indicated *P. aeruginosa* strains grown for 24 hr at 37°C under Tsi6 depletion-inducing (+IPTG) or non-inducing (−IPTG) conditions are shown. The parental strain is the same as in (C).

See also Movies S1 and S2 and Tables S2 and S3.

with the tip of the VgrG spike (Shneider et al., 2013). Despite recent advances in our understanding of the mechanisms underlying T6S-dependent interbacterial interactions, the structure of a T6 effector in complex with a VgrG family protein has remained elusive.

The genome of *Pseudomonas aeruginosa* encodes three T6SSs; each mediate antagonistic interactions with contacting Gram-negative bacterial cells (Hood et al., 2010; Jiang et al., 2014; Russell et al., 2013). The most extensively studied of these is the Hcp secretion island I-encoded T6SS (H1-T6SS), which delivers at least six effectors to recipients. Prior work established that one of these, type VI secretion exported 6 (Tse6), is a predicted transmembrane protein that contains a PAAR repeat domain, is exported in a VgrG-dependent manner, and is active in the cytoplasm of target cells (Whitney et al., 2014). Here, we demonstrate that Tse6 intoxicates by depleting cells of the related co-factors β-nicotinamide adenine dinucleotide (NAD^+) and NAD^+ phosphate ($NADP^+$), thereby simultaneously inhibiting anabolic and catabolic processes required for homeostasis and growth. We make the surprising discovery that Tse6 requires interaction with translation elongation factor Tu for delivery into recipient cells and define the structural and biochemical basis for interbacterial transfer of this membrane-associated toxin.

RESULTS

Tse6 Is a Bacteriostatic Toxin

We previously found that Tse6 is an H1-T6SS-dependent antibacterial effector that requires *vgrG1* for intercellular delivery (Whitney et al., 2014). We further demonstrated that the toxic activity of Tse6 resides in its C terminus and can be neutralized by expression of a cognate immunity protein, Tsi6 (Figures 1A and 1B). Additionally, sequence and structural prediction algorithms identify a PAAR domain ($Tse6_{PAAR}$) flanked by transmembrane segments in the N terminus of the protein.

The toxin domain of Tse6 does not bear homology to characterized proteins. The majority of studied antibacterial T6S effectors act on structures that are important for cellular integrity (Russell et al., 2014). Accordingly, intoxication by these effectors promotes morphological changes and cell lysis (LeRoux et al., 2012). We examined *P. aeruginosa* cells undergoing Tse6-based intoxication via depletion of Tsi6. The H1-T6SS is quiescent in monoculture, thus we performed this and subsequent experiments in a background with activated expression of the system (Δ*retS*) (LeRoux et al., 2015). Single-cell analyses showed Tse6-intoxicated cells displayed a dramatic increase in division time, but generally maintained their structural integrity (Figure 1C; Movies S1 and S2). The markedly slower growth of these cells was also apparent macroscopically; strains depleted of Tsi6 failed to form visible colonies after 24 hr of incubation (Figure 1D). Depletion of Tsi6 from *P. aeruginosa* cells lacking H1-T6SS function (Δ*tssM1*), and thus the capacity to transport effectors intercellularly, yielded indistinguishable effects, indicating that the toxin domain of Tse6 accesses the cytoplasm of donor cells prior to export.

Tse6 Resembles Mono-ADP-Ribosyltransferase Toxins

To gain further insight into Tse6 function, we determined the 1.4 Å resolution crystal structure of its C-terminal toxin domain (residues 282–430, $Tse6_{282\text{-}CT}$) in complex with Tsi6 (Figure 2A;

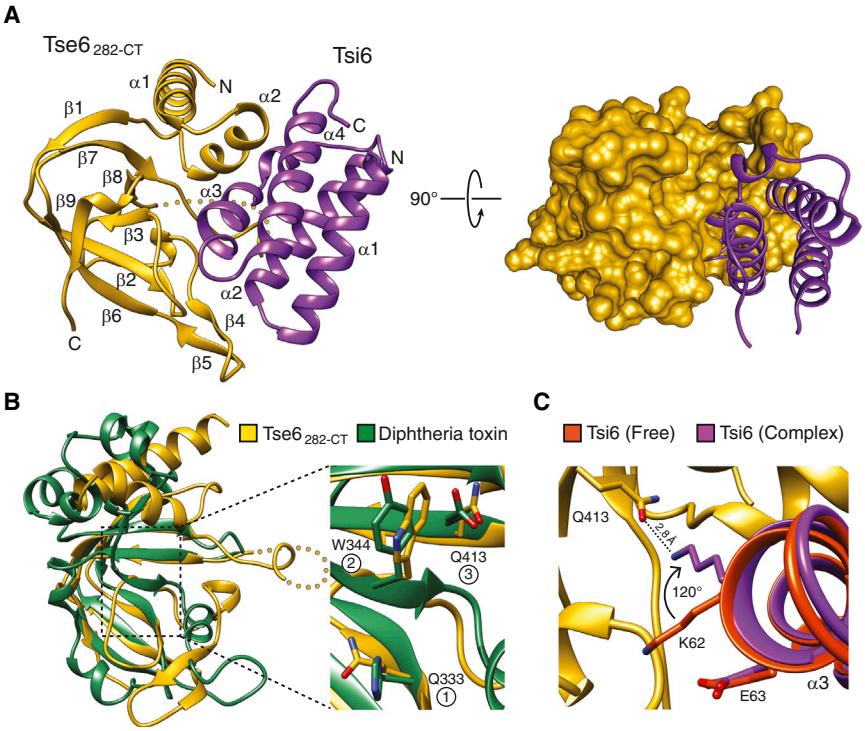


Figure 2. The Toxin Domain of Tse6 Adopts a mART Fold and Harbors a Putative NAD$^+$ Binding Site

(A) Overall structure of the Tse6$_{282\text{-CT}}$–Tsi6 complex. Tse6$_{282\text{-CT}}$ is shown in ribbon (left) and space-filling (right) representations. Secondary structure elements are labeled. Dots denote a disordered segment (amino acids 400-408) of Tse6$_{282\text{-CT}}$ that was not modeled.

(B) Tse6$_{282\text{-CT}}$ resembles mART toxins. Structural alignment of Tse6$_{282\text{-CT}}$ with the catalytic domain of diphtheria toxin (PDB: 4AE1). Inset shows a structural alignment of the three conserved NAD$^+$ binding residues (circled numbers) of diphtheria toxin and Tse6. The numbers correspond to amino acid positions within Tse6.

(C) Tsi6 interacts with the putative NAD$^+$ binding pocket of Tse6. Structural alignment of free Tsi6 and Tsi6 bound to Tse6$_{282\text{-CT}}$. The structure of Tsi6 does not change significantly upon complex formation (e.g., Glu63), except for Lys62, which rotates ~120° and interacts with Gln413 of Tse6.

See also Figure S1 and Tables S1–S3.

Table S1). Importantly, expression of Tse6$_{282\text{-CT}}$ alone induced stasis in *Escherichia coli*, recapitulating the phenotype of the full-length toxin in *P. aeruginosa* (Figure S1A). Tse6$_{282\text{-CT}}$ adopts a mixed α/β fold comprised of two N-terminal α helices and a central core that is formed by two perpendicularly oriented β sheets (Figure 2A). A search of the PDB using DALI indicated that the closest structural homologs of Tse6$_{282\text{-CT}}$ are the catalytic domains of bacterial mono-ADP-ribosyltransferase (mART) toxins (Holm and Rosenström, 2010; Simon et al., 2014). Identified members of this family include diphtheria toxin (DT) from *Corynebacterium diphtheriae* (*Z* score, 4.7; C_α root-mean-square deviation [RMSD] of 4.3 Å over 87 equivalent positions) and Exotoxin A (ExoA) from *P. aeruginosa* (*Z* score, 3.0; C_α RMSD of 2.8 Å over 73 equivalent positions). These and other characterized bacterial mART enzymes are secreted virulence factors that transfer the ADP-ribose moiety of NAD$^+$ onto eukaryotic proteins, typically leading to target protein inactivation, dramatic changes in cellular physiology, and, often, cell death (Simon et al., 2014).

Despite a high degree of sequence divergence within mART proteins, they possess a structurally conserved β sheet core that harbors the molecular determinants for NAD$^+$ binding (Fieldhouse et al., 2010; Zhang et al., 2014). Structural alignment of Tse6$_{282\text{-CT}}$ with DT shows that peripheral secondary structure elements differ significantly, while the two β sheets that comprise the core overlay well (Figure 2B). Characterized mART enzymes are subdivided into two main groups based on the identities of amino acid residues at three positions involved in NAD$^+$ binding (Fieldhouse and Merrill, 2008). Members of the DT group use His, Tyr, and Glu, whereas Cholera toxin-type (CT) proteins retain Glu at position 3, but use Arg and Ser at positions 1 and 2, respectively. Strict conservation of the glutamate between the two groups may reflect its role in stabilizing the oxocarbenium intermediate that forms upon nicotinamide dissociation from ADP-ribose during the catalytic cycle (Yates et al., 2006). Our structure indicates that Tse6 residues differ from those of both mART groups at each position involved in NAD$^+$ binding, including the placement of a non-acidic residue at position 3 (Gln413) (Figure 2B). Nonetheless, the pocket lined by these residues is the principal site of Tsi6 binding, suggesting its importance for the toxic activity of Tse6. In total, our structural analyses suggest that Tse6$_{282\text{-CT}}$ is a mART fold enzyme with unique substrate binding and catalytic motifs.

Tsi6 assumes an all α-helical fold that arranges into a four-helix bundle (Figure 2A). A search of the PDB indicates that Tsi6 shares structural similarity with several proteins of unknown function including Nmul_A1745 from *Nitrosospira multiformis* (*Z* score, 10.4; C_α RMSD of 2.2 Å over 86 equivalent positions) and PA2107 from *P. aeruginosa* (*Z* score, 9.5; C_α RMSD of 2.1 Å over 82 equivalent positions). The Tse6$_{282\text{-CT}}$–Tsi6 interaction involves extensive contacts between α3 of Tsi6 and the putative NAD$^+$ binding pocket of Tse6$_{282\text{-CT}}$. Interface analysis indicates that complex formation between Tse6$_{282\text{-CT}}$ and Tsi6 buries 1,348 Å^2 of solvent-accessible surface area. Isothermal titration calorimetry (ITC) measurements yielded a dissociation constant of 31 nM for the complex (Figure S1B).

To identify the conformational changes within Tsi6 required for inhibition of Tse6 activity, we determined the 1.9 Å crystal structure of Tsi6 in isolation (Table S1). Overall, the structure of free Tsi6 does not differ significantly from that of Tsi6 in complex with Tse6$_{282\text{-CT}}$ (C_α RMSD of 0.4 Å) (Figure S1C). This includes amino acid side chains of Tsi6 involved in the interaction with Tse6$_{282\text{-CT}}$, with the notable exception of Lys62, which rotates approximately 120° around C_β to form a hydrogen bond with the putative NAD$^+$ binding residue at position 3, Gln413 (Figure 2C). Taken together, our structural data suggest that Tsi6

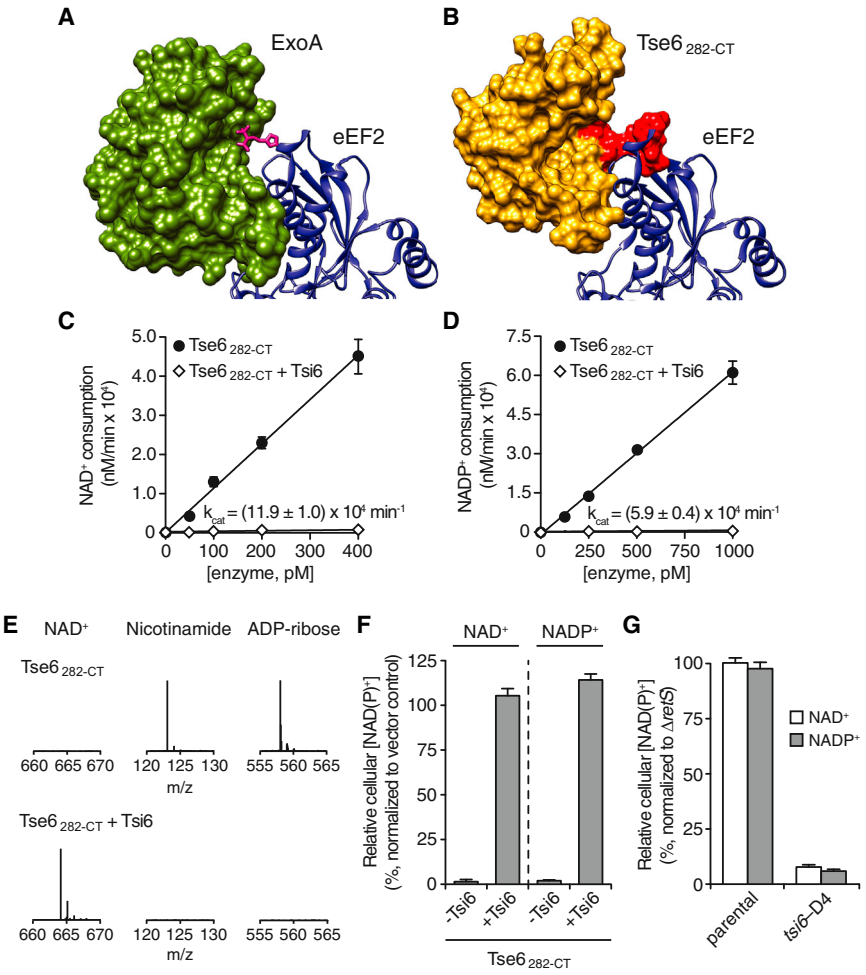


Figure 3. Tse6 Is an NAD(P)$^+$ Glycohydrolase Toxin

(A) mART toxins possess open active sites that allow for docking of their protein targets. Co-crystal structure of *P. aeruginosa* ExoA and eukaryotic elongation factor 2 (eEF2) from Jørgensen et al. (2005). The diphthamide moiety of eEF2 that is ADP-ribosylated by ExoA is shown in pink as a stick representation.

(B) Structural superposition of Tse6$_{282\text{-CT}}$ with ExoA predicts a steric clash with eEF2. The clash occurs through the conserved [K/R]STxxPxxDxx [S/T] motif of Tse6 (red).

(C and D) Tse6$_{282\text{-CT}}$ exhibits NAD(P)$^+$ glycohydrolase activity. Rate of NAD$^+$ (C) and NADP$^+$ (D) consumption by purified Tse6$_{282\text{-CT}}$ in the presence and absence of Tsi6. Each enzyme concentration was assayed in triplicate, and error bars represent ± SD.

(E) Mass spectra of the products generated by Tse6-catalyzed breakdown of NAD$^+$. Peaks corresponding to nicotinamide ([M+H]$^+$, *m/z* = 123.1) and ADP-ribose ([M-H]$^-$, *m/z* = 558.3) were identified in the reaction containing Tse6$_{282\text{-CT}}$, whereas NAD$^+$ ([M+H]$^+$, *m/z* = 664.4) was identified in the reaction containing the Tse6$_{282\text{-CT}}$–Tsi6 complex.

(F) NAD(P)$^+$ levels in *E. coli* cells expressing Tse6$_{282\text{-CT}}$ (−Tsi6) or co-expressing Tse6$_{282\text{-CT}}$ and Tsi6 (+Tsi6) relative to empty vector. Cellular NAD(P)$^+$ levels were assayed 60 min after induction of Tse6$_{282\text{-CT}}$ expression.

(G) Relative NAD(P)$^+$ levels in the indicated *P. aeruginosa* strains 45 min after induction of Tsi6 degradation. Strains correspond to those used in Figure 1C. Error bars represent ± SD (n = 3).

See also Figure S2 and Tables S2 and S3.

inhibits the activity of Tse6 through direct occlusion of its putative NAD$^+$ binding site.

Tse6 Exhibits NAD(P)$^+$ Glycohydrolase Activity

The structure of Tse6$_{282\text{-CT}}$ implies that the toxin may exert its effects within recipient cells via mono-ADP-ribosylation of an unknown bacterial protein. An important feature of characterized mART enzymes that facilitates their transferase activity is an open active site that allows docking of the acceptor protein. This concept is exemplified by the co-crystal structure of *P. aeruginosa* ExoA in complex with eukaryotic elongation factor 2 (Figure 3A) (Jørgensen et al., 2005). However, structural superposition of Tse6$_{282\text{-CT}}$ with ExoA predicts a steric clash between Tse6$_{282\text{-CT}}$ and a proteinaceous ADP-ribose acceptor (Figure 3B). Interestingly, the structural element of Tse6$_{282\text{-CT}}$ that prohibits accommodation of a high-molecular-weight acceptor is comprised of a motif conserved among Tse6 orthologs ([K/R]STxxPxxDxx[S/T]), implying that this region is important for Tse6 function (Zhang et al., 2012). Consistent with these data, incubation of purified Tse6 with *P. aeruginosa* or *E. coli* cell lysates containing ^{32}P-NAD$^+$ did not lead to observable transfer of ^{32}P-ADP-ribose to a protein target (data not shown).

Given the limited accessibility of the Tse6 active site, we hypothesized that the protein might instead function as an NAD$^+$ glycohydrolase. Although it is less common within the mART superfamily of enzymes, NAD$^+$ glycohydrolase activity has been observed for the SPN toxin of *Streptococcus pyogenes* (Ghosh et al., 2010). To test our hypothesis that Tse6 is an NAD$^+$ glycohydrolase enzyme, we performed kinetic analyses of NAD$^+$ consumption by Tse6$_{282\text{-CT}}$. Whereas mART enzymes exhibit only low levels of NAD$^+$ hydrolysis (<10 min^{-1}), we found that purified Tse6$_{282\text{-CT}}$ catalyzes NAD$^+$ breakdown at a rate of approximately 1.2×10^5 min^{-1} (Figure 3C) (Ghosh et al., 2010). This activity was reduced to background by the addition of 1.5 molar equivalents of Tsi6 to the reaction mixture, suggesting NAD$^+$ degradation is a physiologically relevant activity of the toxin. Given the structural similarity between NAD$^+$ and its phosphorylated derivative NADP$^+$, we also tested the ability of Tse6$_{282\text{-CT}}$ to consume NADP$^+$. Breakdown of this dinucleotide occurred at a comparable rate (6.0×10^4 min^{-1}), suggesting that Tse6 degrades both NAD$^+$ and NADP$^+$ (NAD(P)$^+$) (Figure 3D).

Rather than hydrolytically cleaving their substrates, some NAD$^+$-degrading enzymes generate a cyclic product that is a characterized signaling molecule in eukaryotes (Guse, 2000).

The fluorescence assay we employed does not distinguish between cyclized and non-cyclized forms of ADP-ribose, thus we used mass spectrometry (MS) to analyze the reaction products of Tse6 and NAD^+. Nicotinamide and ADP-ribose were the only detectable products, defining Tse6 as an $NAD(P)^+$-glycohydrolase enzyme (Figure 3E).

Tse6 Induces Bacteriostasis by Depleting Cellular $NAD(P)^+$ Levels

Our biochemical data show that Tse6 rapidly hydrolyzes $NAD(P)^+$ in vitro; however, it is possible that we observe this activity due to the absence of an appropriate ADP-ribose acceptor molecule. To address this possibility, we expressed $Tse6_{282\text{-}CT}$ in *E. coli* and measured endogenous $NAD(P)^+$ levels. Upon induction of $Tse6_{282\text{-}CT}$ expression, we found that *E. coli* cells contained vastly reduced cellular concentrations of NAD^+ and $NADP^+$ relative to the vector control (Figure 3F). Co-expression with Tsi6 restored $NAD(P)^+$, indicating that the loss of the dinucleotides is a direct consequence of $Tse6_{282\text{-}CT}$ activity.

Next, we sought to measure the influence of endogenous Tse6 on $NAD(P)^+$ levels in intoxicated *P. aeruginosa* cells. In agreement with our findings in *E. coli*, intracellular intoxication caused by depletion of Tsi6 led to a profound decrease in $NAD(P)^+$ (Figure 3G). The precise measurement of Tse6-catalyzed $NAD(P)^+$ depletion during intercellular intoxication is complicated by high background levels of the dinucleotides derived from donor cells, which as intoxication of recipient cells proceeds, constitute an increasingly large proportion of the total cellular population. To partially overcome this, we examined a time point at which recipient cells have begun to experience intoxication, but are not yet depleted from the population. Comparing total NAD^+ levels in conjunction with donor and recipient colony-forming units (CFU) to a reference experiment, we confirmed a significant reduction in NAD^+ within recipient cells (Figure S2). Based on these findings, we propose that the toxicity elicited by Tse6 is due to depletion of cellular $NAD(P)^+$ pools.

Tse6 Participates in a Five-Protein Complex Containing Elongation Factor Tu

Bioinformatic analyses predict that Tse6 is a PAAR domain-containing integral membrane protein (Figure 1B). To test whether Tse6 resides in membranes, we generated a *P. aeruginosa* strain producing a functional fusion of Tse6 to the vesicular stomatitis virus glycoprotein epitope from the native *tse6* locus (*tse6-V*) (Figure S3A). Western blot analysis of the soluble and membrane fractions of this strain revealed that despite high-confidence prediction of transmembrane domains within Tse6, the majority of the protein is soluble (Figure 4A). Based on structural studies of PAAR domains in complex with VgrG-like chimeras, it has been speculated that effectors containing this domain interact with VgrG proteins (Shneider et al., 2013). We hypothesized that Tse6 could be solubilized within donor cells by virtue of association with VgrG1 via its PAAR domain. Indeed, in the absence of VgrG1, we observed significant repartitioning of Tse6 to the membrane fraction of cells. Tse6 remained soluble in a strain lacking *tssM1*, indicating that the localization of the toxin is not generally sensitive to T6 function.

Motivated by the finding that Tse6 is a soluble protein in the presence of VgrG1, we used co-immunoprecipitation to probe for a physical interaction between the proteins. Surprisingly, this led to the identification of a putative complex containing Tse6, Tsi6, VgrG1, PA0094, and translation elongation factor Tu (EF-Tu) (Figure 4B). PA0094 is a member of a recently described group of effector-specific accessory factors that facilitate delivery of their cognate effectors (Alcoforado and Coulthurst, 2015). Henceforth, we refer to PA0094 as effector-associated gene with *tse6* (EagT6).

The identification of EF-Tu in a complex containing Tse6 was unexpected. This conserved bacterial protein is a GTPase that delivers newly charged aminoacyl-tRNA molecules to the ribosome during translation elongation (Voorhees and Ramakrishnan, 2013). Interactions between T6 effectors and essential bacterial proteins have not been described; therefore, we decided to probe the functional significance of this observation. To test whether Tse6 and EF-Tu interact directly, we initiated experiments to evaluate *P. aeruginosa* EF-Tu ($EF\text{-}Tu^{PA}$) binding to the soluble region of Tse6 ($Tse6_{222\text{-}CT}$) in vitro. Interestingly, during the course of this work, we noted $Tse6_{222\text{-}CT}$ associates with *E. coli* EF-Tu ($EF\text{-}Tu^{EC}$), which is 88% identical to $EF\text{-}Tu^{PA}$. Purification of N-terminal Tse6 truncations narrowed the region responsible for EF-Tu binding to the last 165 amino acids of the toxin (Figure 4C). Since EF-Tu does not bind the glycohydrolase domain of Tse6 (residues 282-CT), we reasoned that the interaction with EF-Tu requires amino acids 265–282 of the toxin. Despite considerable sequence divergence from *P. aeruginosa* Tse6, orthologs of the toxin from *P. putida* and *P. syringae* also co-purified with $EF\text{-}Tu^{EC}$ (Figures S3B and S3C).

Next, we measured the binding affinity of $Tse6_{265\text{-}CT}$ to $EF\text{-}Tu^{EC}$ and $EF\text{-}Tu^{PA}$ using ITC. Guanosine triphosphate (GTP) hydrolysis by EF-Tu is coupled to significant structural changes in the protein; therefore, we investigated both GTP- and GDP- (EF-Tu•GDP) bound conformations of the molecule (Clark and Nyborg, 1997). In line with our purification results, we found that $Tse6_{265\text{-}CT}$ interacts tightly with both $EF\text{-}Tu^{EC}$•GDP (K_d = 81 nM) and $EF\text{-}Tu^{PA}$•GDP (K_d = 23 nM) (Figures S3D and S3E). Application of the antibiotic Aurodox, which locks EF-Tu into its GTP-bound conformation, reduced the affinity for Tse6 by approximately 10-fold (Figure S3F) (Vogeley et al., 2001). In summary, these data indicate that Tse6 binds directly to the GDP form of EF-Tu within a larger, multiprotein complex.

Structure of the Tse6-EF-Tu Complex

Though all cells require $NAD(P)^+$, the process of translation does not rely on these co-factors. Thus, the significance of Tse6 interaction with EF-Tu was not apparent. As a first step toward defining the relevance of the Tse6-EF-Tu complex, we determined the 3.5 Å crystal structure of $Tse6_{265\text{-}CT}$ bound to $EF\text{-}Tu^{PA}$•GDP (Figure 4D; Table S1).

Overall, the structure of $Tse6_{265\text{-}CT}$ is highly similar to $Tse6_{282\text{-}CT}$ (C_α RMSD of 0.7 Å). The most striking divergence between the two structures is the ordering and 60° hinge-like movement of the [K/R]STxxPxxDxx[S/T] motif-containing loop, henceforth referred to as the Tse6 activation loop. This results in a ~15-Å displacement of Asp396 that directs its side chain into the putative $NAD(P)^+$ binding site (Figure 4E). Asp396 is the sole invariant

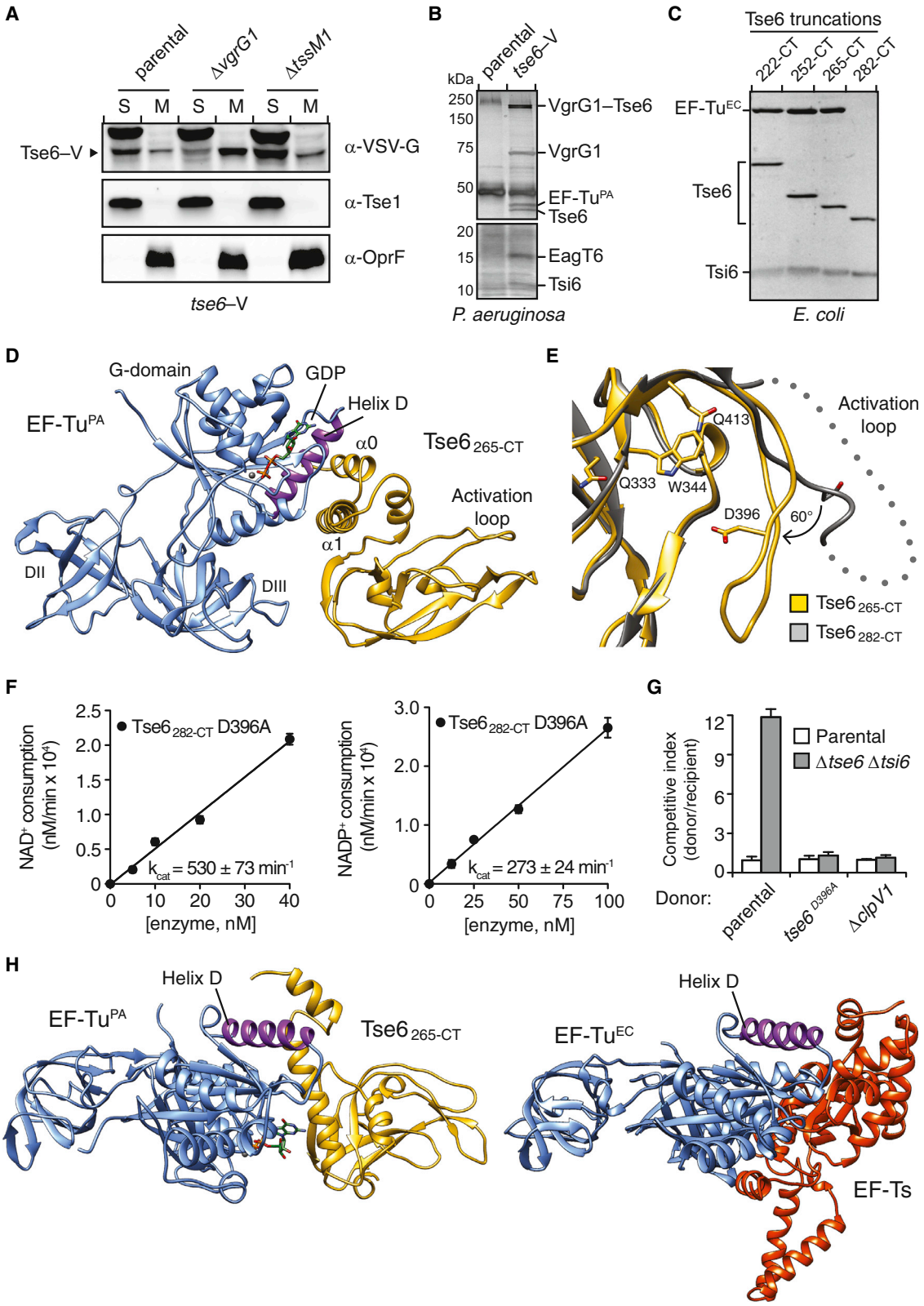


Figure 4. Tse6 Participates in a Multi-protein Complex and Binds Helix D of EF-Tu through Residues N-Terminal to Its Toxin Domain

(A) Tse6 is a membrane protein that is solubilized by VgrG1. Western blot analysis of the soluble (S) and membrane (M) fractions of the indicated *P. aeruginosa* strains. Tse1 and OprF serve as soluble and membrane controls, respectively.

(legend continued on next page)

acidic residue among Tse6 orthologous proteins, leading us to postulate that it serves a role analogous to the conserved glutamic acid at position 3 within the DT and CT mART families (Figure S4) (Zhang et al., 2012). Consistent with this hypothesis, purified $Tse6_{282\text{-}CT}^{D396A}$ displayed approximately 225-fold reduced $NAD(P)^+$ glycohydrolase activity relative to the wild-type protein, and a *P. aeruginosa* strain producing $Tse6^{D396A}$ from the native *tse6* locus did not exhibit Tse6-based intercellular intoxication (Figures 4F and 4G).

In accordance with our Tse6 truncation studies, interaction with $EF\text{-}Tu^{PA}$ is mediated by residues immediately N-terminal to the toxin domain (residues 265–291). This basic segment forms two α helices ($\alpha0$ and $\alpha1$) that engage in numerous salt bridges with acidic side chains on the GTPase domain (G domain) of $EF\text{-}Tu^{PA}$. Interestingly, the $EF\text{-}Tu^{PA}$ residues involved in this interaction are found on helix D, which functions as the key interaction site for both the guanine exchange factor EF-Ts and the ribosome (Figure 4H) (Kawashima et al., 1996).

Interaction with EF-Tu Is Required for the Delivery of Tse6 into Recipient Cells

Our structure of $Tse6_{265\text{-}CT}$–$EF\text{-}Tu^{PA}$ shows that a spatially confined cluster of electrostatic interactions facilitates binding of the proteins (Figures 5A and 5B). We reasoned that this interaction mechanism might afford an opportunity to dissect the functional significance of the interaction via site-directed mutagenesis. A non-conserved leucine residue (Leu270) was identified within the patch of basic amino acids on $\alpha0$ that mediate EF-Tu binding (Figure 5C). We postulated that an acidic residue substituted at this position would disrupt charge complementarity between the proteins. As predicted, $Tse6_{222\text{-}CT}^{L270E}$ did not co-purify with $EF\text{-}Tu^{EC}$, whereas variants containing a more conservative substitution at this site (L270A) or an analogous substitution on the opposite face of $\alpha0$ (A268E) retained $EF\text{-}Tu^{EC}$ binding (Figure 5D).

Encouraged by our in vitro data, we next generated a *P. aeruginosa* strain expressing $Tse6^{L270E}$-V from the native *tse6* locus. An immunoprecipitation and growth competition experiment utilizing this strain showed that $Tse6^{L270E}$-V displays a specific defect in EF-Tu interaction and is unable to intoxicate recipient cells (Figures 5E and 5F). We conclude that association with EF-Tu is essential for Tse6-based toxicity.

Tse6-based intercellular intoxication can be viewed as a number of discrete processes. We considered the involvement in and requirement for EF-Tu in (1) the stability of Tse6, (2) the enzymatic activity of Tse6, (3) Tse6 export from donor cells, and (4) entry of Tse6 into recipient cells. Since $Tse6^{L270E}$ is present at equal concentrations as the wild-type protein, we ruled out a requirement for EF-Tu in Tse6 stability. Our biochemical data show that the catalytic domain of Tse6 degrades NAD^+ rapidly, at a rate consistent with a known cytotoxic NAD^+ glycohydrolase enzyme (Ghosh et al., 2010). Therefore, one possibility is that residues N-terminal to the toxin domain are auto-inhibitory and that EF-Tu binding to this region relieves this inhibition. Indeed, the activation loop of Tse6 differs significantly in position between the $Tse6_{282\text{-}CT}$–Tsi6 and $Tse6_{265\text{-}CT}$–$EF\text{-}Tu^{PA}$ structures, suggesting that either EF-Tu induces a conformational change in the toxin or immunity protein binding excludes this loop from the active site (Figure 4E). We found that a purified Tse6 variant that includes the EF-Tu binding region ($Tse6_{222\text{-}CT}$) catalyzes NAD^+ hydrolysis at a rate indistinguishable to the toxin domain alone (Figure 5G). Furthermore, the activity of this protein was unaffected by the addition of excess $EF\text{-}Tu^{PA}$.

Next, we considered the possibility that association of Tse6 with EF-Tu is required for export of the toxin from donor cells. However, we found that both cellular and extracellular levels of $Tse6^{L270E}$-V are similar to the wild-type protein (Figure S5). These experiments further showed that unlike the Hcp-associated effector Tse1, Tse6 accumulation in the exo-proteome of *P. aeruginosa* is only partially dependent on H1-T6SS function. The significance of this is not yet understood; however, strains lacking the T6S ATPase ClpV1 or the core integral membrane protein TssM1 yielded similar results.

Since interaction with EF-Tu is dispensable for Tse6 catalytic activity and export, we deduced that interaction with the translation factor must be required for Tse6 to reach the cytoplasm of recipient cells. In further support of this contention, we found $Tse6^{L270E}$-V is as active in intracellular intoxication triggered by Tsi6 depletion as the wild-type protein (Figure 5H). Together with our findings that $Tse6^{L270E}$-V is incapable of Tse6-based intercellular intoxication despite its unencumbered transit of the T6SS, these data indicate that interaction with EF-Tu grants Tse6 access to the cytoplasm of recipient cells.

Ultrastructure of a T6 Effector-VgrG Complex

VgrG is thought to serve as the T6S protein that pierces the outer membrane of recipient cells, granting its bound cognate effector(s) access to the periplasm of target cells (Silverman et al.,

(B) Silver-stained SDS-PAGE analysis of proteins enriched by anti-VSV-G immunoprecipitation from *P. aeruginosa* strains encoding Tse6 (parental) and Tse6-V. The labels indicate the identities of proteins that specifically co-precipitate with Tse6-V as determined by MS. In addition to their monomeric forms, VgrG1 and Tse6 form a high-molecular-weight complex that is resistant to heat and SDS denaturation.

(C) A 17-amino-acid segment of Tse6 mediates interaction with EF-Tu. Coomassie-stained SDS-PAGE analysis of purified Tse6 truncations. All truncations were expressed with Tsi6 and assessed for co-purification with endogenous $EF\text{-}Tu^{EC}$.

(D) Overall structure of the $Tse6_{265\text{-}CT}$–$EF\text{-}Tu^{PA}$ complex. Secondary structure elements involved in the interaction are labeled.

(E) The Tse6 activation loop harbors Asp396 and rotates toward the active site of Tse6 in the $Tse6_{265\text{-}CT}$–$EF\text{-}Tu^{PA}$ structure relative to its position in the $Tse6_{282\text{-}CT}$–Tsi6 structure. Dots denote a disordered segment (amino acids 400-408) of $Tse6_{282\text{-}CT}$ that was not modeled.

(F) $Tse6_{282\text{-}CT}$ D396A exhibits significantly reduced $NAD(P)^+$ glycohydrolase activity. Rate of NAD^+ (left) and $NADP^+$ (right) consumption by purified $Tse6_{282\text{-}CT}$ D396A.

(G) Asp396 is critical for Tse6-based intercellular toxicity. Growth competition experiments between the indicated *P. aeruginosa* donor and recipient strains. Donor and recipient strains were mixed 1:1, grown for 24 hr on solid media, and differentiated using blue/white screening.

(H) Helix D of EF-Tu is the site of interaction for both $Tse6_{265\text{-}CT}$ (left) and the guanine exchange factor EF-Ts (right). In all panels, error bars represent ± SD (n = 3). See also Figures S3 and S4and Tables S1–S3.

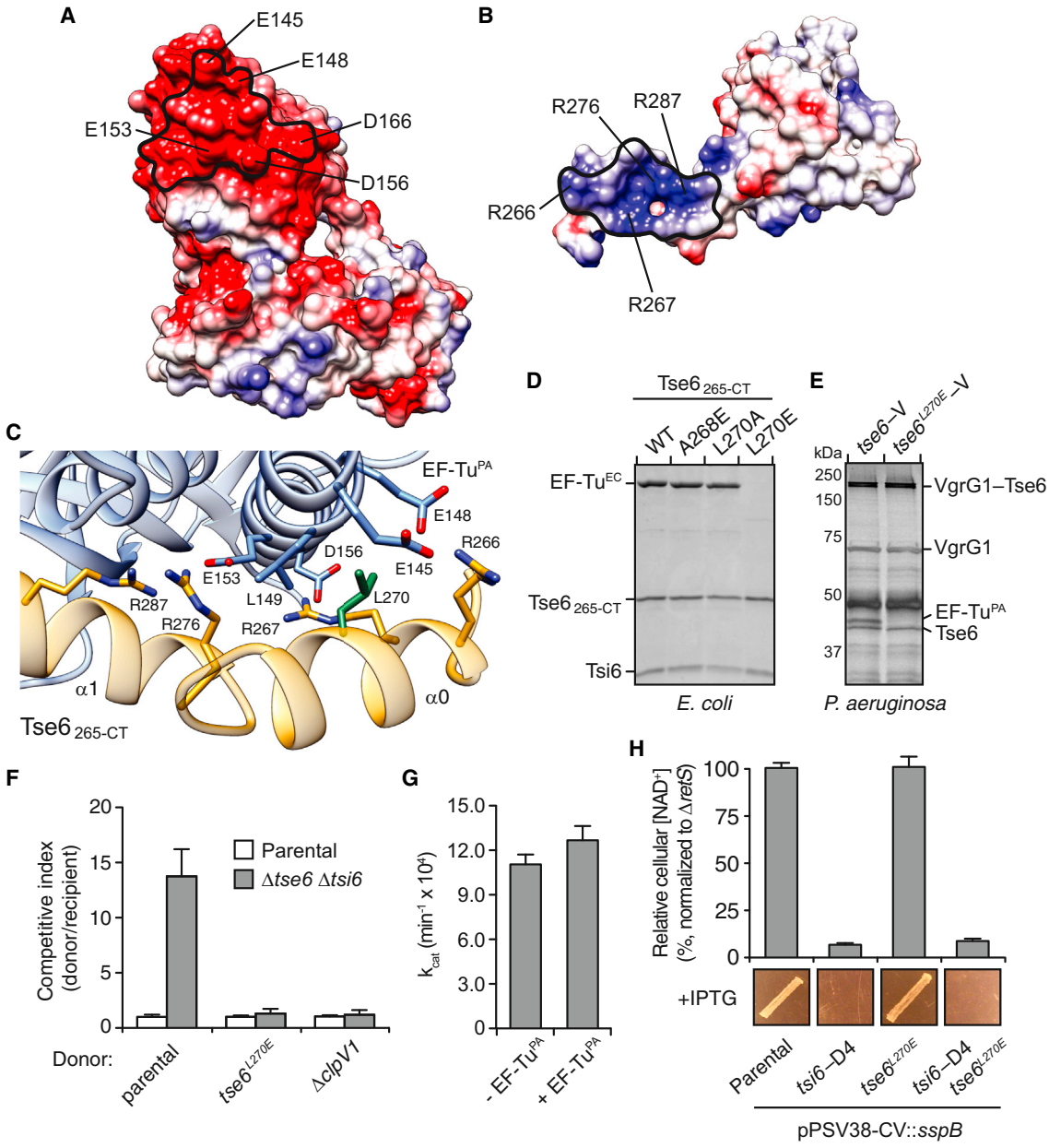


Figure 5. Interaction with EF-Tu Is Required for Tse6-Based Intercellular Toxicity

(A and B) An electrostatic patch mediates interaction between EF-Tu and $Tse6_{265\text{-}CT}$. Electrostatic surface representation of $EF\text{-}Tu^{PA}$ (A) and $Tse6_{265\text{-}CT}$ (B). Residues participating in the interaction are labeled and outlined in black.

(C) Close-up view of the $Tse6_{265\text{-}CT}$–$EF\text{-}Tu^{PA}$ interaction. Secondary structure elements referred to in the text are labeled.

(D and E) An L270E variant of Tse6 does not interact with $EF\text{-}Tu^{EC}$ or $EF\text{-}Tu^{PA}$. (D) Coomassie-stained SDS-PAGE analysis of purified $Tse6_{265\text{-}CT}$ variants. All variants were expressed with Tsi6 and assessed for their ability to co-purify with endogenous $EF\text{-}Tu^{EC}$. (E) Silver-stained SDS-PAGE analysis of proteins enriched by anti-VSV-G immunoprecipitation from *P. aeruginosa* strains encoding Tse6-V (parental) and Tse6-V^{L270E}. Enriched low-molecular-weight proteins (EagT6 and Tsi6) not shown.

(F) Tse6 requires interaction with EF-Tu to intoxicate recipient cells. Outcome of growth competition experiments between the indicated *P. aeruginosa* donor strains and a parental (Δ*retS*) or Tse6-susceptible (Δ*tse6* Δ*tsi6*) recipient. The competitive index is calculated as the change (final/initial) in ratio of donor to recipient CFU.

(G) Interaction with $EF\text{-}Tu^{PA}$ does not enhance NAD^+ glycohydrolase activity of $Tse6_{222\text{-}CT}$. Reactions were performed using 500 pM $Tse6_{222\text{-}CT}$ in the presence or absence of 1 μM $EF\text{-}Tu^{PA}$.

(H) Interaction with EF-Tu is not required for Tse6-based intracellular intoxication. NAD^+ levels in the indicated *P. aeruginosa* strains 45 min after induction of Tsi6 degradation (top). Patches of the indicated *P. aeruginosa* strains grown for 24 hr at 37°C under Tsi6 depletion-inducing (+IPTG) conditions (bottom). The parental strain is the same as in Figure 1C. Error bars represent ± SD (n = 3).

See also Figure S5and Tables S2 and S3.

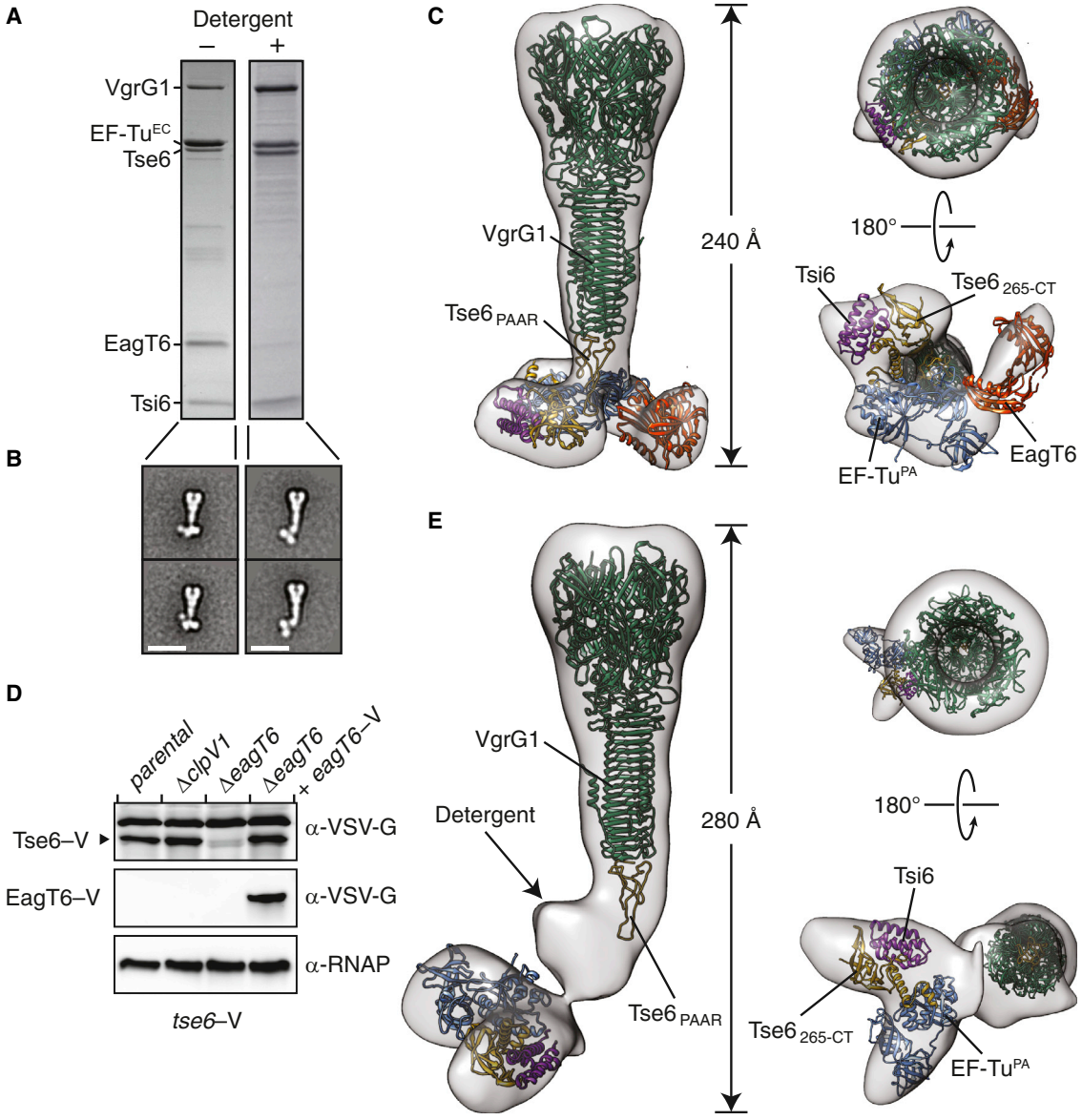


Figure 6. Two Conformations of the Tse6 Secretory Particle Revealed by Electron Microscopy

(A and B) Addition of detergent dissociates EagT6 from the Tse6 secretion particle and causes a conformational change. (A) Coomassie-stained SDS-PAGE analysis and (B) representative class averages of purified Tse6-containing complex in the presence and absence of 0.03% β-D-dodecylmaltopyranoside.

(C) 3D density map and molecular fitting of the Tse6-Tsi6-VgrG1-EagT6-EF-Tu^{PA} complex. The identity of each subunit is indicated. The model for Tse6$_{PAAR}$ was generated using *Phyre* (Kelley and Sternberg, 2009).

(D) Tse6 requires EagT6 for intracellular accumulation. Western blot analysis of Tse6 levels in the indicated *P. aeruginosa* strains. RNA polymerase (RNAP) is used as a loading control.

(E) 3D density map and molecular fitting of the detergent-bound Tse6–Tsi6–VgrG1–EF-Tu^{PA} complex. Scale bars, 20 nm.

See also Figures S6 and S7and Tables S2 and S3.

2012). The PAAR domain of effectors associates with the tip of VgrG; however, the placement of additional effector domains, as well as accessory proteins, in this particle is not known (Shneider et al., 2013). To gain insight into the topology of an effector-loaded VgrG complex, we examined purified C-terminally octa-histidine-tagged Tse6 in complex with Tsi6, VgrG1, EagT6, and EF-Tu^{EC} using negative-stain electron microscopy (EM) (Figures 6A and 6B, left).

Analysis of 12,000 single particles permitted the calculation of a 3D map of the complex resolved to 22 Å (Figures 6C and S6A–S6D). VgrG proteins have a characteristic structure that was readily apparent within the map (Shneider et al., 2013). Fortuitously, the unpublished X-ray crystal structure of *P. aeruginosa* VgrG1 is available in the PDB (PDB: 4MTK); the location of this trimeric assembly in our structure was unambiguous. For estimating the placement of Tse6, Tsi6, and EF-Tu, we were able

to utilize our assorted high-resolution structures of these proteins to produce a ternary complex that largely conformed to regions of density found near the end of the complex predicted to initiate contact with recipient cells. An additional constraint on the location of Tse6 is its PAAR domain, which could be modeled with high confidence bound to the tip of VgrG1 (Figure 6C). The 91 residues connecting $Tse6_{PAAR}$ to the first residue included in our modeled ternary complex (residue 265) are predicted to form a transmembrane helix, followed by a disordered glycine-rich span (not modeled). Ni-NTA-nanogold labeling of the C-terminal His_8-tag of Tse6 provided support for our placement of this linchpin protein within the complex (Figure S6E).

The positions of VgrG, Tse6, Tsi6, and $EF\text{-}Tu^{PA}$ left a protruding region of density surrounding the PAAR domain of Tse6 unoccupied. We postulated that this region of the map corresponds to EagT6. The structure of EagT6 was determined in a high-throughput X-ray crystallography project and made publicly available. The domain-swapped homodimeric protein bears a distinctive horseshoe configuration that we placed in this unoccupied density (Figures 6C and S7A). In this configuration, EagT6 would be predicted to bind $Tse6_{PAAR}$ and its buttressing hydrophobic segments. To test this prediction, we performed co-purification experiments of the proteins in *E. coli*. In contrast to the complex isolated with full-length toxin, $Tse6_{222\text{-}CT}$ did not co-purify with EagT6 (Figure S7B). Although the X-ray crystal structure of EagT6 does not fully agree with the calculated map in this region, these data provide biochemical support for our placement of EagT6 in proximity to the N-terminal domains of Tse6.

To garner additional insight into EagT6 function, we probed the capacity of *P. aeruginosa* Δ*eagT6* to intoxicate Tse6-sensitive recipient cells. This strain failed to elicit Tse6-based toxicity, but retained the capacity to intoxicate recipients using another H1-T6S effector (Figure S7C). EagT6 associates with a region of Tse6 rich in transmembrane domains, suggesting that the accessory factor could act as a chaperone for the toxin. As predicted for substrate-chaperone systems, we found that accumulation of Tse6 is markedly diminished by the absence of EagT6 (Figure 6D). Genetic complementation of this phenotype was achieved with ectopic expression of *eagT6*. Taken together with the findings of Alcoforado and Coulthurst (2015) pertaining to a EagT6-related protein in *Serratia*, we propose that EagT6 functions as a Tse6-specific chaperone.

Tse6 is a transmembrane protein; thus, its transport between cells likely requires shielding of its hydrophobic domains. Based on its orientation and position relative to Tse6 in our model, we posited that EagT6 might chaperone Tse6 by shielding its hydrophobic segments from aqueous mediums during intercellular transport. In support of this hypothesis, EagT6 dissociates from the effector complex in the presence of detergent (Figure 6A, right). To gain structural insight into the consequence of EagT6 release, we examined the Tse6 particle depleted of this protein by negative-stain EM and single-particle reconstruction (Figure 6B, right). Analysis of 11,000 particles permitted the calculation of a 3D map of the complex resolved to 19 Å (Figures 6E and S6F–S6J). Remarkably, we observed a ~40 Å movement of the EF-Tu–$Tse6_{265\text{-}CT}$–Tsi6 sub-complex from VgrG-$Tse6_{PAAR}$. The localization of $Tse6_{265\text{-}CT}$ within the displaced density was verified by Ni-NTA-nanogold labeling (Figure S6K). Accompanying this reorganization, we observed a region of unoccupied density in proximity to the predicted site of the hydrophobic segments of Tse6. Given the capacity of detergent to compete for EagT6 binding to the complex, we hypothesize that ordered detergent molecules bound to the hydrophobic domains of Tse6 occupy this density.

The significance of Tsi6 and EF-Tu within the effector complex is not understood. The toxicity conferred by depletion of Tsi6 from donor cells shows that the catalytic domain of Tse6 is present in the cytoplasm and is in complex with Tsi6 prior to export by the H1-T6SS. Therefore, EF-Tu also likely interacts with the toxin prior to export. If these proteins do not dissociate from the toxin during transit, the structures we obtained would represent the secreted complex. Alternatively, Tsi6 and EF-Tu could be removed during secretion and re-engage the complex in the cytoplasm of recipient cells. In this scenario, the complex we isolated would represent that found in recipient cells with immunity. In total, our ultrastructural analyses of the Tse6 secretory particle define the architecture of an effector-loaded VgrG and suggest a mechanism for deployment of a membrane-associated toxin (Figure 7).

DISCUSSION

We have discovered that Tse6 intoxicates recipient cells by catalyzing the hydrolytic removal of the nicotinamide moiety from NAD^+ and $NADP^+$. This mechanism has not been described for an interbacterial toxin, but it is consistent with the general observation that T6 effectors act on target molecules that are both essential and highly conserved among bacteria (Russell et al., 2014). The consequence of $NAD(P)^+$ degradation by Tse6 is stasis in most cells, rather than death. The relative benefit(s) of inhibiting the growth of target cells is not yet understood; however, the H1-T6SS Tse2 toxin also induces stasis in recipients (Li et al., 2012). In instances of self-intoxication, the exchange of bacteriostatic toxins could promote the formation of persister cells. A non-mutually exclusive possibility is that when delivered within an effector cocktail, bacteriostatic and bacteriocidal toxins act synergistically.

The requirement for $NAD(P)^+$ extends to all forms of life, raising the possibility that Tse6, and related proteins exported by the T6SS, could intoxicate archaeal and eukaryotic cells. The SPN toxin of *S. pyogenes* provides precedent for the action of a bacterial NAD^+ glycohydrolase against a eukaryotic target, although this is a structurally distinct toxin that utilizes pores introduced by Streptolysin O to gain entry into host cells (Madden et al., 2001; Smith et al., 2011). The H3-T6SS of *P. aeruginosa* has been shown to deliver a phospholipase D toxin to both bacterial and eukaryotic cells, implying that there is not a fundamental barrier to interdomain targeting of effectors by this bacterium (Jiang et al., 2014).

To our knowledge, the requirement for a housekeeping protein in the function of a T6S effector has not previously been observed. Likely owing to the central role of EF-Tu in translation, *tufA*, which encodes EF-Tu, is a slowly evolving bacterial gene (Lathe and Bork, 2001). Thus, if the role of the Tse6 interaction with a cellular housekeeping protein is to grant the toxin access to recipient cells as our data suggest, EF-Tu would allow the toxin to target phylogenetically diverse bacteria. The high concentration of EF-Tu within cells could also contribute to a wide

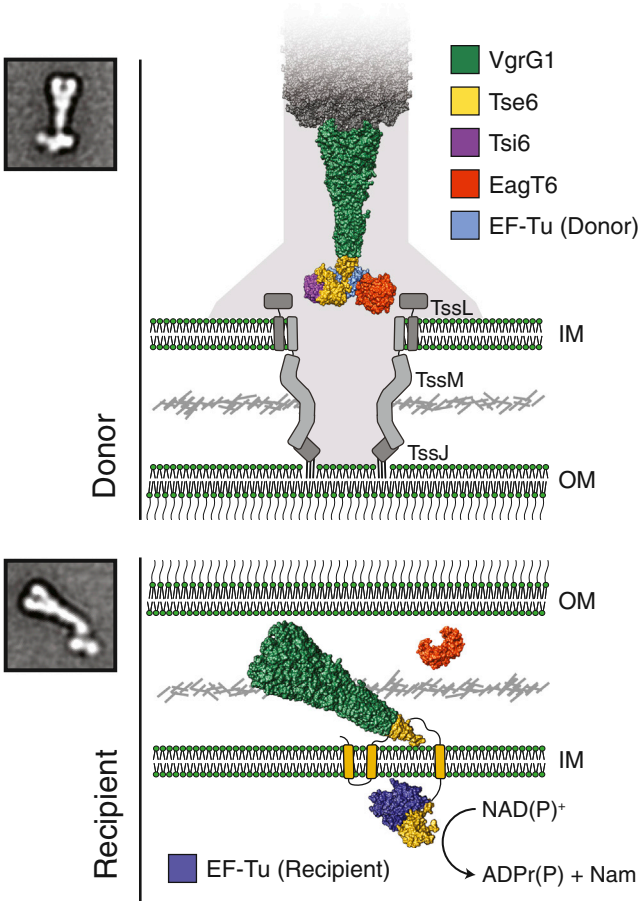


Figure 7. Proposed Model for Tse6 Transport by the T6S Apparatus

The configuration of the Tse6 particle subunits in donor and recipient cells represent those determined in this study in the absence and presence of detergent, respectively. A representative EM class average for each is provided for reference. In the donor cell, Tse6 and associated proteins are bound to the T6S sheath complex (gray; PDB: 3J9O) (Clemens et al., 2015). The T6 trans-envelope complex (composed of TssL, M, and J) is schematized to approximate its recently determined EM structure (Durand et al., 2015), whereas the T6 baseplate-like assembly is depicted in filled gray. In the model, donor cell EF-Tu (light blue) and Tsi6 disengage from the Tse6 secretion particle upon export from the donor cell. Upon crossing the outer membrane (OM) of the recipient cell, EagT6 dissociation frees the hydrophobic domains of Tse6 (yellow rectangles) for incorporation into the recipient inner membrane (IM). Recipient cell EF-Tu (dark blue) facilitates transfer of the $NAD(P)^+$ glycohydrolase domain into the recipient cell cytoplasm by an unknown mechanism. Several other possibilities consistent with all available data are not presented. Most notably, donor-cell-derived EF-Tu may be exported as part of the secretion particle and facilitate Tse6 delivery into recipient cells, or donor-cell-derived EF-Tu may be excluded from the secretion complex.
See also Tables S2 and S3

target range for Tse6, as there would be more tolerance for weakened association between the two proteins driven by EF-Tu sequence divergence.

Although the site of Tse6 binding to EF-Tu would preclude binding of the translation factor to EF-Ts, it is unlikely that Tse6 affects translation (Kawashima et al., 1996). We find Tse6 present at low levels in *P. aeruginosa*; therefore, the yet lower levels within recipient cells would not sequester a functionally significant portion of EF-Tu. Indeed, there is growing evidence that the large pool of EF-Tu is exploited for multiple purposes within bacteria, including *P. aeruginosa* (Balasubramanian et al., 2008; Barel et al., 2008; Defeu Soufo et al., 2010; Kunert et al., 2007; Mohan et al., 2014). Barbier et al. (2013) have found that EF-Tu^{PA} is posttranslationally modified by trimethylation at Lys5. These authors also found that this form of the protein localizes to the cell surface, where it mediates interactions with airway epithelial cells. Whether EF-Tu is actively secreted to the cell surface or if its presence there is a consequence of cell lysis was not determined. It is worth noting that the high concentration of EF-Tu present in culture supernatants through T6-independent mechanisms precluded measurement of the contribution of T6 to EF-Tu export in our study.

Given the changing chemical and physical environments that necessarily accompany translocation across multiple membranes, it is without doubt that effectors delivered intercellularly assume multiple states en route. We have captured just two of these for a T6S effector. Tse6 is the hub of a multi-protein complex in our structures; however, our biochemical data show that it need not interact with any of these proteins in order to catalyze $NAD(P)^+$ degradation. This leaves many open questions, including how does EF-Tu facilitate Tse6 translocation into recipient cells? From our current data, we cannot determine whether EF-Tu derived from donor cells, recipient cells, or both is critical for Tse6 activity. One appealing model consistent with our data holds that Tse6 is delivered to the target cell periplasm, whereupon EagT6 is released and the exposed transmembrane segments of the toxin spontaneously insert into the inner membrane. At this point, translocation of residues N-terminal to the toxin domain and ensuing EF-Tu-binding could serve as a molecular ratchet that favors passage of the remaining toxin domain into the cytoplasm. Interestingly, the EF-Tu binding domain of Tse6 is rich in basic residues, a property of many known cell-penetrating peptides (Bechara and Sagan, 2013).

While this study provides two snapshots of interbacterial protein transport, it also highlights the challenges in understanding this intricate, multi-step process. The Tse6 particle we describe may provide a tractable system for the characterization of additional secretory intermediates. A complete understanding of toxin entry into recipient cells could define novel routes for the delivery of antimicrobials.

EXPERIMENTAL PROCEDURES

Bacterial Strains, Plasmids, and Growth Conditions

All *P. aeruginosa* strains generated were derived from the sequenced strain PAO1 (Stover et al., 2000). *P. aeruginosa* mutants and chromosomal fusions were generated by allelic exchange as described previously (Hood et al., 2010). *E. coli* strains DH5α, BL21(DE3) pLysS, and SM10 were used for plasmid maintenance, gene expression, and conjugative transfer, respectively. A detailed list of strains and plasmids used in this study can be found in Tables S2 and S3.

Crystallization and Structure Determination

Details for the crystallization of $Tse6_{282\text{-}CT}$–Tsi6, Tsi6, and $Tse6_{265\text{-}CT}$–EF-Tu are described in the Supplemental Experimental Procedures. The structures of $Tse6_{282\text{-}CT}$–Tsi6 and Tsi6 were solved by Se-SAD. The $Tse6_{265\text{-}CT}$-EF-Tu structure was solved by molecular replacement using $EF\text{-}Tu^{EC}$•GDP (PDB: 1EFC) as a search model. Details for structure determination and model

refinement are described in the Supplemental Experimental Procedures (Table S1).

Biochemical Assays

Hydrolysis rates of $NAD(P)^+$ by Tse6 were measured using a fluorescence endpoint assay as described previously (Johnson and Morrison, 1970). Determination of relative NAD^+ and $NADP^+$ levels from cell lysates was performed using the NAD/NADH-Glo and NADP/NADPH-Glo bioluminescence assays, respectively, as per the instructions of the manufacturer (Promega). Details can be found in the Supplemental Experimental Procedures.

Bacterial Competition Assays

Intraspecific competition assays between *P. aeruginosa* strains were performed as previously described (Whitney et al., 2014). Briefly, overnight cultures of *P. aeruginosa* strains were mixed in a 1:1 ratio and spotted onto 0.2-μm nitrocellulose membranes overlaid on a 3% agar Luria broth no-salt plate. Competitive indices were calculated by enumerating donor/recipient CFU after 24 hr of growth at 37°C. All competitive indices were adjusted by the donor/recipient ratio in the initial inoculum. Recipient strains express the *lacZ* gene from a neutral phage attachment site to enable their differentiation from unlabeled donor via blue/white screening.

Electron Microscopy and Image Analysis

Protein samples were negatively stained with uranyl formate (SPI Supplies/Structure Probe) and imaged using a JEOL1400 microscope equipped with a LaB_6 cathode operated at 120 kV. Images were recorded at a magnification of 50,000× on a 4k × 4k CMOS camera F416 (TVIPS). Data analysis and further processing was done in SPARX (Hohn et al., 2007). Details can be found in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession numbers for the atomic coordinates of $Tse6_{282\text{-}CT}$—Tsi6, Tsi6, and $Tse6_{265\text{-}CT}$—$EF\text{-}Tu^{PA}$ are PDB: 4ZV0, 4ZUY, and 4ZV4, respectively. The accession numbers for the negative-stain EM maps of the detergent-bound and detergent-free Tse6 secretion particle are EMDB: EMD-3112 and EMD-3113, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, three tables, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.027.

ACKNOWLEDGMENTS

The authors would like to thank H. Kulasekara for assistance with membrane fractionation, C. Outten for providing pRSFDuet-1, W. Catterall for use of the ITC instrument, J. Woodward for assistance with radioactivity experiments and for providing Aurodox, C. Ralston for assistance with X-ray data collection, I. Attree for providing α-OprF antibody, C. Gatsogiannis for electron microscopy expertise, and S. Dove, C. Goulding, C. Hayes, D. Low, A. Merz, D. Veesler, and members of the S.R. and J.D.M. laboratories for helpful discussions. This work was supported by grants from the NIH (AI080609) (to J.D.M.) and by the University of Maryland Baltimore, School of Pharmacy Mass Spectrometry Center (SOP1841-IQB2014) (to D.R.G.). J.C.W. was supported by a postdoctoral research fellowship from the Canadian Institutes of Health Research, D.Q. was supported by a Chemiefonds fellowship from the Fonds der Chemischen Industrie, S.S. was supported by a Mary Gates Research Scholarship, and J.D.M. holds an Investigator in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund.

Received: May 18, 2015
Revised: July 15, 2015
Accepted: August 19, 2015
Published: October 8, 2015

REFERENCES

Alcoforado, D.J., and Coulthurst, S.J. (2015). Intra-species competition in Serratia marcescens is mediated by type VI secretion Rhs effectors and a conserved effector-associated accessory protein. J. Bacteriol. *197*, 2350–2350.

Balasubramanian, S., Kannan, T.R., and Baseman, J.B. (2008). The surface-exposed carboxyl region of Mycoplasma pneumoniae elongation factor Tu interacts with fibronectin. Infect. Immun. *76*, 3116–3123.

Barbier, M., Owings, J.P., Martínez-Ramos, I., Damron, F.H., Gomila, R., Blázquez, J., Goldberg, J.B., and Albertí, S. (2013). Lysine trimethylation of EF-Tu mimics platelet-activating factor to initiate Pseudomonas aeruginosa pneumonia. MBio *4*, e00207–e00213.

Barel, M., Hovanessian, A.G., Meibom, K., Briand, J.P., Dupuis, M., and Charbit, A. (2008). A novel receptor - ligand pathway for entry of Francisella tularensis in monocyte-like THP-1 cells: interaction between surface nucleolin and bacterial elongation factor Tu. BMC Microbiol. *8*, 145.

Basler, M., Pilhofer, M., Henderson, G.P., Jensen, G.J., and Mekalanos, J.J. (2012). Type VI secretion requires a dynamic contractile phage tail-like structure. Nature *483*, 182–186.

Bechara, C., and Sagan, S. (2013). Cell-penetrating peptides: 20 years later, where do we stand? FEBS Lett. *587*, 1693–1702.

Clark, B.F., and Nyborg, J. (1997). The ternary complex of EF-Tu and its role in protein biosynthesis. Curr. Opin. Struct. Biol. *7*, 110–116.

Clemens, D.L., Ge, P., Lee, B.Y., Horwitz, M.A., and Zhou, Z.H. (2015). Atomic structure of T6SS reveals interlaced array essential to function. Cell *160*, 940–951.

Defeu Soufo, H.J., Reimold, C., Linne, U., Knust, T., Gescher, J., and Graumann, P.L. (2010). Bacterial translation elongation factor EF-Tu interacts and colocalizes with actin-like MreB protein. Proc. Natl. Acad. Sci. USA *107*, 3163–3168.

Durand, E., Nguyen, V.S., Zoued, A., Logger, L., Péhau-Arnaudet, G., Aschtgen, M.S., Spinelli, S., Desmyter, A., Bardiaux, B., Dujeancourt, A., et al. (2015). Biogenesis and structure of a type VI secretion membrane core complex. Nature *523*, 555–560.

Fieldhouse, R.J., and Merrill, A.R. (2008). Needle in the haystack: structure-based toxin discovery. Trends Biochem. Sci. *33*, 546–556.

Fieldhouse, R.J., Turgeon, Z., White, D., and Merrill, A.R. (2010). Cholera- and anthrax-like toxins are among several new ADP-ribosyltransferases. PLoS Comput. Biol. *6*, e1001029.

Fritsch, M.J., Trunk, K., Diniz, J.A., Guo, M., Trost, M., and Coulthurst, S.J. (2013). Proteomic identification of novel secreted antibacterial toxins of the Serratia marcescens type VI secretion system. Mol. Cell. Proteomics *12*, 2735–2749.

Ghosh, J., Anderson, P.J., Chandrasekaran, S., and Caparon, M.G. (2010). Characterization of Streptococcus pyogenes beta-NAD+ glycohydrolase: re-evaluation of enzymatic properties associated with pathogenesis. J. Biol. Chem. *285*, 5683–5694.

Guse, A.H. (2000). Cyclic ADP-ribose. J. Mol. Med. *78*, 26–35.

Hachani, A., Allsopp, L.P., Oduko, Y., and Filloux, A. (2014). The VgrG proteins are "à la carte" delivery systems for bacterial type VI effectors. J. Biol. Chem. *289*, 17872–17884.

Hohn, M., Tang, G., Goodyear, G., Baldwin, P.R., Huang, Z., Penczek, P.A., Yang, C., Glaeser, R.M., Adams, P.D., and Ludtke, S.J. (2007). SPARX, a new environment for Cryo-EM image processing. J. Struct. Biol. *157*, 47–55.

Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545–W549.

Hood, R.D., Singh, P., Hsu, F., Güvener, T., Carl, M.A., Trinidad, R.R., Silverman, J.M., Ohlson, B.B., Hicks, K.G., Plemel, R.L., et al. (2010). A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria. Cell Host Microbe *7*, 25–37.

Jiang, F., Waterfield, N.R., Yang, J., Yang, G., and Jin, Q. (2014). A Pseudomonas aeruginosa type VI secretion phospholipase D effector targets both prokaryotic and eukaryotic cells. Cell Host Microbe *15*, 600–610.

Johnson, S.L., and Morrison, D.L. (1970). The alkaline reaction of nicotinamide adenine dinucleotide, a new transient intermediate. J. Biol. Chem. *245*, 4519–4524.

Jørgensen, R., Merrill, A.R., Yates, S.P., Marquez, V.E., Schwan, A.L., Boesen, T., and Andersen, G.R. (2005). Exotoxin A-eEF2 complex structure indicates ADP ribosylation by ribosome mimicry. Nature *436*, 979–984.

Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., and Leberman, R. (1996). The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 A resolution. Nature *379*, 511–518.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. *4*, 363–371.

Kunert, A., Losse, J., Gruszin, C., Hühn, M., Kaendler, K., Mikkat, S., Volke, D., Hoffmann, R., Jokiranta, T.S., Seeberger, H., et al. (2007). Immune evasion of the human pathogen Pseudomonas aeruginosa: elongation factor Tuf is a factor H and plasminogen binding protein. J. Immunol. *179*, 2979–2988.

Lathe, W.C., 3rd, and Bork, P. (2001). Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Lett. *502*, 113–116.

LeRoux, M., De Leon, J.A., Kuwada, N.J., Russell, A.B., Pinto-Santini, D., Hood, R.D., Agnello, D.M., Robertson, S.M., Wiggins, P.A., and Mougous, J.D. (2012). Quantitative single-cell characterization of bacterial interactions reveals type VI secretion is a double-edged sword. Proc. Natl. Acad. Sci. USA *109*, 19804–19809.

LeRoux, M., Kirkpatrick, R.L., Montauti, E.I., Tran, B.Q., Peterson, S.B., Harding, B.N., Whitney, J.C., Russell, A.B., Traxler, B., Goo, Y.A., et al. (2015). Kin cell lysis is a danger signal that activates antibacterial pathways of Pseudomonas aeruginosa. eLife *4*, 4.

Li, M., Le Trong, I., Carl, M.A., Larson, E.T., Chou, S., De Leon, J.A., Dove, S.L., Stenkamp, R.E., and Mougous, J.D. (2012). Structural basis for type VI secretion effector recognition by a cognate immunity protein. PLoS Pathog. *8*, e1002613.

Ma, L.S., Hachani, A., Lin, J.S., Filloux, A., and Lai, E.M. (2014). Agrobacterium tumefaciens deploys a superfamily of type VI secretion DNase effectors as weapons for interbacterial competition in planta. Cell Host Microbe *16*, 94–104.

Madden, J.C., Ruiz, N., and Caparon, M. (2001). Cytolysin-mediated translocation (CMT): a functional equivalent of type III secretion in gram-positive bacteria. Cell *104*, 143–152.

Mohan, S., Hertweck, C., Dudda, A., Hammerschmidt, S., Skerka, C., Hallström, T., and Zipfel, P.F. (2014). Tuf of Streptococcus pneumoniae is a surface displayed human complement regulator binding protein. Mol. Immunol. *62*, 249–264.

Russell, A.B., Hood, R.D., Bui, N.K., LeRoux, M., Vollmer, W., and Mougous, J.D. (2011). Type VI secretion delivers bacteriolytic effectors to target cells. Nature *475*, 343–347.

Russell, A.B., LeRoux, M., Hathazi, K., Agnello, D.M., Ishikawa, T., Wiggins, P.A., Wai, S.N., and Mougous, J.D. (2013). Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors. Nature *496*, 508–512.

Russell, A.B., Peterson, S.B., and Mougous, J.D. (2014). Type VI secretion system effectors: poisons with a purpose. Nat. Rev. Microbiol. *12*, 137–148.

Shneider, M.M., Buth, S.A., Ho, B.T., Basler, M., Mekalanos, J.J., and Leiman, P.G. (2013). PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. Nature *500*, 350–353.

Silverman, J.M., Brunet, Y.R., Cascales, E., and Mougous, J.D. (2012). Structure and regulation of the type VI secretion system. Annu. Rev. Microbiol. *66*, 453–472.

Silverman, J.M., Agnello, D.M., Zheng, H., Andrews, B.T., Li, M., Catalano, C.E., Gonen, T., and Mougous, J.D. (2013). Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates. Mol. Cell *51*, 584–593.

Simon, N.C., Aktories, K., and Barbieri, J.T. (2014). Novel bacterial ADP-ribosylating toxins: structure and function. Nat. Rev. Microbiol. *12*, 599–611.

Smith, C.L., Ghosh, J., Elam, J.S., Pinkner, J.S., Hultgren, S.J., Caparon, M.G., and Ellenberger, T. (2011). Structural basis of Streptococcus pyogenes immunity to its NAD+ glycohydrolase toxin. Structure *19*, 192–202.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature *406*, 959–964.

Vogeley, L., Palm, G.J., Mesters, J.R., and Hilgenfeld, R. (2001). Conformational change of elongation factor Tu (EF-Tu) induced by antibiotic binding. Crystal structure of the complex between EF-Tu.GDP and aurodox. J. Biol. Chem. *276*, 17149–17155.

Voorhees, R.M., and Ramakrishnan, V. (2013). Structural basis of the translational elongation cycle. Annu. Rev. Biochem. *82*, 203–236.

Whitney, J.C., Beck, C.M., Goo, Y.A., Russell, A.B., Harding, B.N., De Leon, J.A., Cunningham, D.A., Tran, B.Q., Low, D.A., Goodlett, D.R., et al. (2014). Genetically distinct pathways guide effector export through the type VI secretion system. Mol. Microbiol. *92*, 529–542.

Yates, S.P., Jørgensen, R., Andersen, G.R., and Merrill, A.R. (2006). Stealth and mimicry by deadly bacterial toxins. Trends Biochem. Sci. *31*, 123–133.

Zhang, D., de Souza, R.F., Anantharaman, V., Iyer, L.M., and Aravind, L. (2012). Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol. Direct *7*, 18.

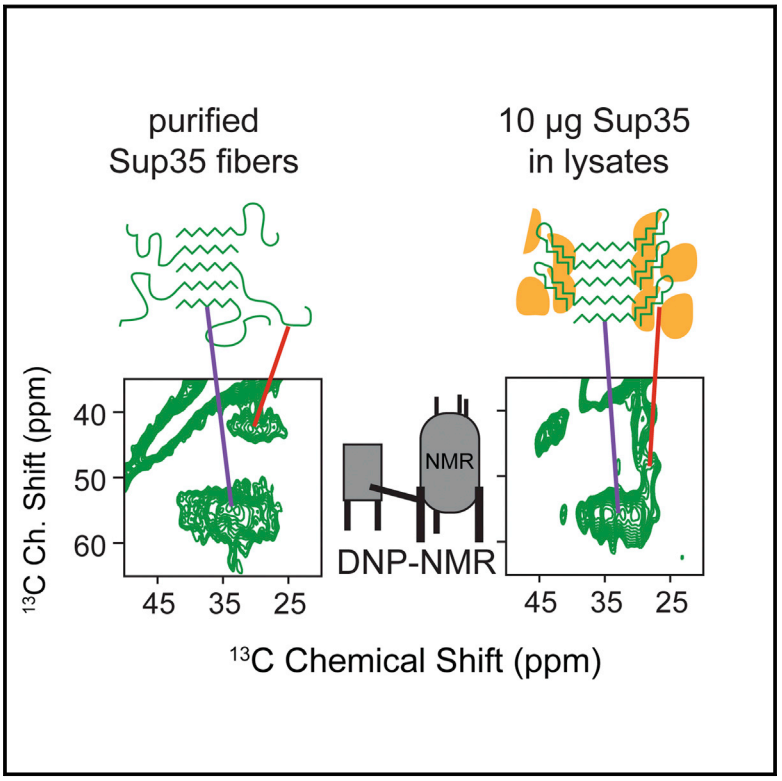
Zhang, D., Iyer, L.M., Burroughs, A.M., and Aravind, L. (2014). Resilience of biochemical activity in protein domains in the face of structural divergence. Curr. Opin. Struct. Biol. *26*, 92–103.

Article

Cell

Sensitivity-Enhanced NMR Reveals Alterations in Protein Structure by Cellular Milieus

Graphical Abstract



Highlights

- Sensitivity-enhanced DNP NMR enables detection of proteins at endogenous levels
- DNP NMR allows specific detection of a protein in complex physiological environments
- Cellular environments alter the structure of intrinsically disordered regions

Authors

Kendra K. Frederick, Vladimir K. Michaelis, Björn Corzilius, Ta-Chung Ong, Angela C. Jacavone, Robert G. Griffin, Susan Lindquist

Correspondence

kendra.frederick@utsouthwestern.edu (K.K.F.),
lindquist_admin@wi.mit.edu (S.L.)

In Brief

Sensitivity-enhanced NMR enabling structural analysis of protein at endogenous levels in a native biological context reveals that the cellular environment alters the structure of an intrinsically disordered protein domain.

Frederick et al., 2015, Cell *163*, 1–9
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.024

CellPress

Please cite this article in press as: Frederick et al., Sensitivity-Enhanced NMR Reveals Alterations in Protein Structure by Cellular Milieus, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.024

Article

Cell

Sensitivity-Enhanced NMR Reveals Alterations in Protein Structure by Cellular Milieus

Kendra K. Frederick,[1,5,*] Vladimir K. Michaelis,[2] Björn Corzilius,[2,6] Ta-Chung Ong,[2,7] Angela C. Jacavone,[2] Robert G. Griffin,[2,4] and Susan Lindquist[1,3,4,*]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Department of Chemistry and Francis Bitter Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Co-senior author
[5]Present address: Department of Biophysics, UT Southwestern, Dallas, TX 75390, USA
[6]Present address: Institute of Physical and Theoretical Chemistry, Institute of Biophysical Chemistry, and Center for Biomolecular Magnetic Resonance, Goethe-University, Frankfurt am Main, Germany
[7]Present address: Department of Chemistry, Laboratory of Inorganic Chemistry, ETH-Zürich, CH-8093 Zürich, Switzerland
*Correspondence: kendra.frederick@utsouthwestern.edu (K.K.F.), lindquist_admin@wi.mit.edu (S.L.)
http://dx.doi.org/10.1016/j.cell.2015.09.024

SUMMARY

Biological processes occur in complex environments containing a myriad of potential interactors. Unfortunately, limitations on the sensitivity of biophysical techniques normally restrict structural investigations to purified systems, at concentrations that are orders of magnitude above endogenous levels. Dynamic nuclear polarization (DNP) can dramatically enhance the sensitivity of nuclear magnetic resonance (NMR) spectroscopy and enable structural studies in biologically complex environments. Here, we applied DNP NMR to investigate the structure of a protein containing both an environmentally sensitive folding pathway and an intrinsically disordered region, the yeast prion protein Sup35. We added an exogenously prepared isotopically labeled protein to deuterated lysates, rendering the biological environment "invisible" and enabling highly efficient polarization transfer for DNP. In this environment, structural changes occurred in a region known to influence biological activity but intrinsically disordered in purified samples. Thus, DNP makes structural studies of proteins at endogenous levels in biological contexts possible, and such contexts can influence protein structure.

INTRODUCTION

Structural investigations of biomolecules are typically confined to in vitro systems under limited conditions. Although investigations yield invaluable insights, such experiments can never capture all aspects of complex biological environments. Proteins must fold into their active conformations in complex environments. This situation becomes perilous when considering proteins that must attain a particular conformation, but whose energetic folding landscapes are rather flat or have several local minima. In these cases, the environment can clearly influence the conformation by favoring one pathway over another. Such decisions can have striking biological consequences, as is the case for a variety of protein folding diseases (Dobson, 2001). The effect of environment becomes even more critical when considering the substantial fraction of the human proteome that encodes disordered proteins (Dunker et al., 2001). Intrinsically disordered proteins (IDPs) are important components of the cellular signaling machinery, allowing the same polypeptide to undertake different interactions with different consequences (Wright and Dyson, 2015). Yet, structural characterization of these domains is notoriously difficult (Uversky, 2013).

Yeast prions present both of these structural challenges as they have both environmentally sensitive protein folding landscapes as well as intrinsically disordered regions. Yeast prions have provided a paradigm shift in our understanding of heritable biological information. They allow specific biological traits to be encoded and inherited solely though self-templating protein conformations. When a protein switches to its prion conformation, its function changes. This altered function is passed from generation to generation by conformational self-templating and catalyzed division of the template to daughter cells. The most extensively studied yeast prion, [*PSI*$^+$] (Cox, 1965), is an amyloid conformer of the translation termination factor Sup35. In purified amyloid fibrils of the prion domain of Sup35, called NM, the N-terminal domain (N) adopts a beta-sheet-rich amyloid conformation while the adjacent middle domain (M) is intrinsically disordered (Frederick et al., 2014; Krishnan and Lindquist, 2005; Luckgei et al., 2013; Toyama et al., 2007). However, this is unlikely to be the case in vivo: the M domain is known to interact with many other biomolecules, including protein remodeling factors that regulate prion inheritance. As a consequence, mutations in the M domain (Helsen and Glover, 2012; Liu et al., 2002), or changes in the levels of protein chaperones (e.g., Hsp70) and protein remodeling factors (e.g., Hsp104) (Kiktev et al., 2012; Masison et al., 2009; Tuite et al., 2011) have profound effects on prion propagation. NM also physically associates with protein chaperones (Allen et al., 2005), and at least one chaperone binding site has been localized to the M domain of NM (Helsen and Glover, 2012). Finally, a host of genetic data

suggests that protein-based inheritance is sensitive to the combination and stoichiometry of many other proteins, meaning that isolated study of prion structure can offer at best only partial insight into this paradigm shifting biology.

Interest in prions is highlighted by the fact that similar structural transitions figure in the pathologies of a wide variety of human diseases. Prion strains were first described for the mammalian prion protein, PrP (Chien et al., 2004; Prusiner et al., 1998) and polymorphic amyloid forms have been reported for a variety of amyloidogenic protein associated with neurodegenerative disease (Guo et al., 2013; Kodali et al., 2010; Nekooki-Machida et al., 2009; Petkova et al., 2005). Upon structural characterization, only a portion of the protein is sequestered into the amyloid core. The amyloid cores of these fibers are flanked by intrinsically disordered regions (Heise et al., 2005; Helmus et al., 2008; Wasmer et al., 2009). More recently, amyloid forms of such proteins were demonstrated to have prion-like self-templating dispersion properties in vivo (Jucker and Walker, 2013; Polymenidou and Cleveland, 2012; Watts et al., 2013).

Nuclear magnetic resonance (NMR) is a powerful spectroscopic method for studying molecular structure and dynamics. A key strength of this technique is that it can be used to study non-crystalline, amorphous samples. Indeed, there have been a handful of high-profile in-cell NMR studies (Banci et al., 2013; Freedberg and Selenko, 2014; Inomata et al., 2009; Reckel et al., 2012; Sakakibara et al., 2009; Selenko et al., 2006; Vaiphei et al., 2011). These studies suggest that while protein structure can be perturbed, it is largely unchanged by the cellular context. However, these studies employed solution-state NMR to detect proteins at concentrations two or more orders of magnitude above endogenous levels inside cells, radically altering endogenous stoichiometries. Because solution-state NMR is limited by molecular tumbling times (that depend upon molecular size and solvent viscosity), the minority of the protein that might interact with cellular components would likely be undetectable. Moreover, because this population would comprise a small fraction of the total biomolecule, it would be difficult, if not impossible, to detect the resulting signal loss. Solid state NMR is not limited by molecular correlation times in this way. Instead, solid-state NMR is limited by its low sensitivity. Dynamic nuclear polarization (DNP) has the potential to alleviate this limitation by dramatically increasing the sensitivity of NMR spectroscopy, through the transfer of the large spin polarization that is associated with unpaired electrons to nearby nuclei (Abragam, 1983; Slichter, 1990). Theoretically, DNP can reduce experimental times by more than five orders of magnitude; an experiment that would require decades without DNP can be collected in a day with DNP. However, just as for other structural biology techniques, DNP sensitivity enhancements are critically dependent on experimental conditions (Ni et al., 2013) and sample composition (Akbey et al., 2010, 2013; Takahashi et al., 2014) and the specificity of NMR is critically dependent upon the choice of isotopic labeling (Wang et al., 2013). There is growing interest in application of DNP to complex systems. Several groups have applied DNP to investigate membrane proteins that were overexpressed to high levels in bacteria and have directly examined both concentrated membrane fractions and whole cells (Jacso et al., 2012; Renault et al., 2012; Yamamoto et al., 2015). We report conditions that enable high polarization transfer efficiencies in biologically complex environments. These are large enough to allow the characterization of a single protein at endogenous concentrations in its native environment. Structural methods to investigate either intrinsically disordered proteins or environmentally sensitive protein folding are limited. Here, we present a generalizable approach for investigation of both of these challenging structural puzzles that lie at the heart of both fundamental biological questions and human diseases. Moreover, we demonstrate that including the biological context can influence protein structure.

RESULTS

NM Adopts an Amyloid Form in Cell Lysates at Low Concentrations

We first confirmed the NM protein adopted its active conformation at endogenous concentrations in a native environment. Previous studies have employed extensive serial dilution and propagation in purified in vitro conditions (Frederick et al., 2014). To ensure that the exogenously added protein was faithfully templated by the prion conformers from the cell lysate, we probed its structural state using semi-denaturing detergent agarose gel electrophoresis (SDD-AGE) (Bagriantsev et al., 2006; Halfmann and Lindquist, 2008). NM did not form amyloid in lysates from cells that do not harbor the [*PSI*$^{+}$] prion form of Sup35 (Figure 1B). In contrast, NM was templated into an amyloid form by both purified pre-formed fibers and lysates from cells that harbored the prion. We determined the concentration of templated, exogenously added NM was ~1 μM by immunoblot (Figure 1C), in good agreement with previously reported endogenous Sup35 concentrations of 2.5–5 μM (Ghaemmaghami et al., 2003). In this way, we prepared samples of isotopically labeled NM amyloids at endogenous levels in a complex biological environment.

Sensitivity and Specificity of DNP Magic Angle Spinning NMR

Having established that NM adopts an amyloid conformation in cellular lysates, we prepared recombinant, ^{1}H, ^{13}C-labeled NM and added it to cell lysates that had been grown in deuterated media with carbon isotopes in natural abundance. This created a spectroscopically active prion protein in an NMR silent cellular background. We prepared the sample for DNP magic angle spinning (MAS) NMR by addition of cryoprotectant (glycerol) and a stable biradical TOTAPOL (Song et al., 2006). We collected 1D ^{13}C{^{1}H} cross polarization (CP) spectra of the cellular lysates both with and without microwave-driven polarization transfer from electrons to nuclei (DNP). Experiments using DNP resulted in significant signal enhancements relative to conventional NMR. DNP signal enhancements (ε) at 211 MHz were between 50- and 115-fold (Figures 2 and S1). The carbonyl carbon enhancements were similar to the maximal enhancements obtained for the reference system proline (ε = 130) for this instrumental configuration. This establishes that DNP MAS NMR is well-suited to study complex biological mixtures.

DNP enhances the NMR signal of all ^{13}C atoms in the sample. Interestingly, in samples with the uniformly ^{13}C-labeled protein at

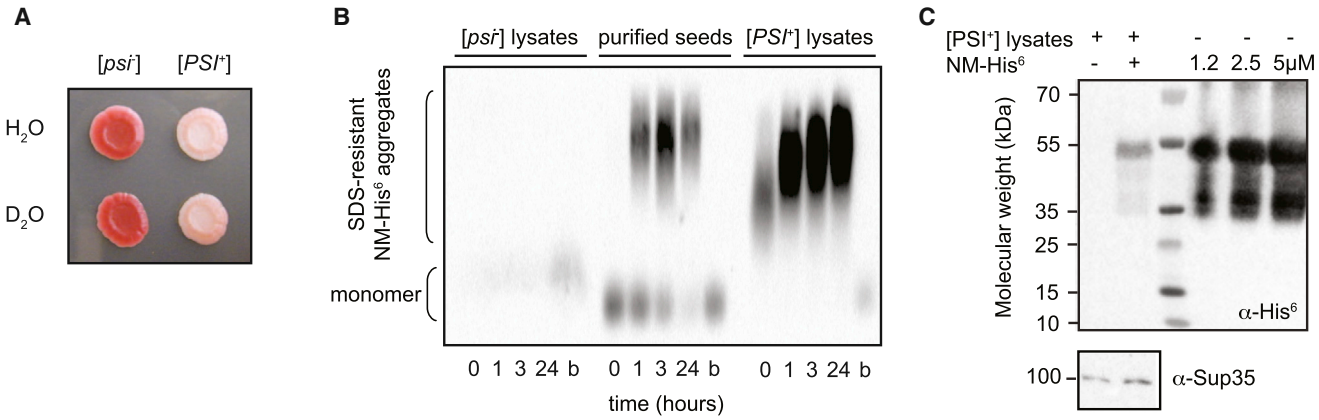


Figure 1. NM Adopts an Amyloid Form in Cell Lysates at Low Concentrations

(A) Prion status is maintained for yeast grown on deuterated media, indicating that the [*PSI*$^+$] protein folding phenotypes were robust to growth in a deuterated environment. Phenotypically prion minus [*psi*$^-$] (red) or prion plus [*PSI*$^+$] (pink) yeasts were grown to mid-log phase in media made with either H_2O or D_2O and then spotted onto a one-fourth YPD plate. Plates were incubated at 30°C for 1 week.

(B) Amyloid formation of purified NM-His^6 was visualized by SDD-AGE using an anti-His^6 antibody in prion minus ([*psi*$^-$]) cell lysates that do not contain endogenous prion templates, in the presence of 2% (w/w) purified amyloid seeds and in prion containing ([*PSI*$^+$]) cellular lysates that contain endogenous prion templates. As with the endogenous prion, boiling (b) destroyed the templated amyloid aggregates.

(C) NM templated into the amyloid form in yeast cell lysates is not degraded and is present at endogenous levels. Full-length endogenous Sup35 runs at 100 kDa and is visualized with an antibody specific to the C-terminal domain. Cellular lysates both with and without exogenously added NM-His^6 as well as a concentration gradient of purified NM-His^6 were boiled in 2% SDS to denature any higher order aggregates and separated by SDS-PAGE before western blotting with an antibody specific for the His^6 epitope. NM-His^6 runs at 55 kDa. The endogenous concentration of Sup35 is between 2.5 and 5 μM. The ECL signal for NM-His^6 in lysates is less intense than that of purified NM-His^6 at a concentration of 1.2 μM, indicating that the concentration of exogenously added NM in the NMR sample is below 1.2 μM.

endogenous concentrations, the ^{13}C content from its natural abundance is an order of magnitude larger than that of added NM protein. However, because the natural abundance of ^{13}C is 1.1%, only 0.01% of the ^{13}C sites in the cell lysate were adjacent to another ^{13}C site. Conversely, all the ^{13}C sites in the exogenously added uniformly ^{13}C-labeled NM had adjacent ^{13}C sites. To isolate ^{13}C signals from NM and filter out background ^{13}C signals from the cell lysates, we collected one-bond ^{13}C-^{13}C dipolar recoupled correlation spectra using proton driven spin diffusion (PDSD) (Szeverenyi et al., 1982). In this 2D experiment, on-diagonal peaks report on all ^{13}C sites in the sample while off-diagonal peaks, or cross-peaks, report only on ^{13}C sites that are directly bonded to another ^{13}C site. To determine the contributions of cell lysates to the ^{13}C-^{13}C correlation spectra, we used signals from β1,3-glucan, a major cell wall component that is well-resolved from protein signals. As expected, the ratio of the cross-peak C_1-C_2 signal intensity relative to the diagonal C_1 signal for β1,3-glucan was 2.5% ± 2% of that for yeast grown on uniformly ^{13}C-enriched glucose. However, the ratio of the protein carbonyl carbon (C′)–carbon alpha (C_α) cross-peak signal intensity relative to the diagonal C′ signal intensity for the protein backbone region was 10-fold higher (21% ± 2%) for the natural abundance sample containing added ^{13}C NM than the ratio for the β1,3-glucan region. The protein signal was an order of magnitude larger than the lysate background expected from natural abundance, establishing that the cross-peak signals in the ^{13}C-^{13}C correlation spectra report on the added NM and not on ^{13}C in the cellular lysates. To completely eliminate any concerns about the contribution of natural abundance ^{13}C from the cellular lysates, samples of prion-containing yeasts for structural investigations were grown with ^{13}C-depleted (99.9% ^{12}C) glucose as the carbon source, further reducing the ^{13}C cross-peak intensity from the cellular lysates by two orders of magnitude. Thus, the combination of DNP with this isotopic labeling scheme provides the sensitivity and specificity to observe a protein at endogenous levels in a biologically complex native environment.

To investigate the structural influence of cellular lysates on NM amyloid assembly, we compared spectra of NM fibers at endogenous levels in cellular lysates to spectra of purified lysate-templated NM fibers (Frederick et al., 2014). We conducted these experiments at higher magnetic fields (700 MHz rather than 211 MHz) to achieve significant improvements in spectral resolution (Barnes et al., 2012; Michaelis et al., 2014). We collected a one-bond ^{13}C-^{13}C dipolar-assisted rotational resonance (DARR) (Takegoshi et al., 2001) correlation spectrum on 1 mg of cryoprotected, purified NM fibrils in 6 hr. For 10 μg of NM fibrils in unlabeled cellular lysates, we collected a one-bond ^{13}C-^{13}C DARR spectrum for 1 week. As expected, no cross-peaks for β1,3 glucan were present in spectra of cellular lysates grown in depleted ^{13}C glucose. Inhomogeneous line broadening due to experimental temperatures required for DNP (83 K) potentially counteracts any gain in spectral resolution from higher magnetic fields. Thus, we compared spectra of purified NM fibers under DNP conditions to spectra of purified NM fibers at room temperature. In both samples, most of the resonances overlapped due to the number of sites and highly degenerate amino acid composition of this protein, a common feature of prion proteins (Frederick et al., 2014). Nonetheless, the line widths of isolated side chain sites in the DNP spectra at 83 K are similar to those of

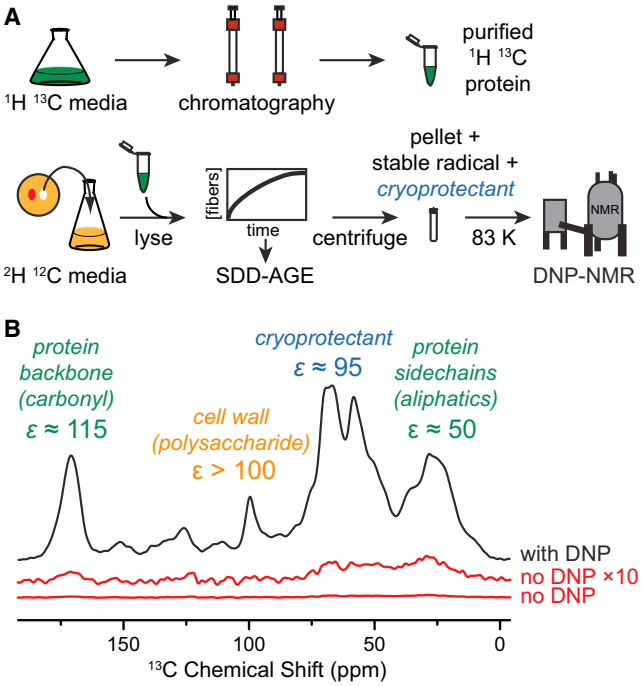


Figure 2. Dynamic Nuclear Polarization Enhances NMR Signals in Cellular Lysates

(A) Preparation of samples for DNP MAS NMR of proteins at endogenous levels in biological environments. The protein of interest is expressed on isotopically enriched media and purified. The cellular background comes from cells grown on media containing D_2O. The cells are lysed and the isotopically labeled protein is added exogenously to whole lysate. The mixture is pelleted, the pellet is resuspended in a matrix containing stable radical and cryoprotectant, and the mixture is frozen for analysis by DNP MAS NMR.

(B) One-dimensional $^{13}C\{^{1}H\}$ spectra both with (black) and without DNP enhancement by microwaves (red). Dynamic nuclear polarization gave large signal enhancements (ε) for uniformly ^{1}H ^{13}C-labeled NM in a deuterated matrix of cellular lysates containing a 60:30:10 (v/v) mixture of d_8-glycerol:D_2O:H_2O and 10 mM TOTAPOL at 211 MHz/140 GHz with $\omega/2\pi$ = 4.3 kHz and a sample temperature of 83 K.

See also Figure S1.

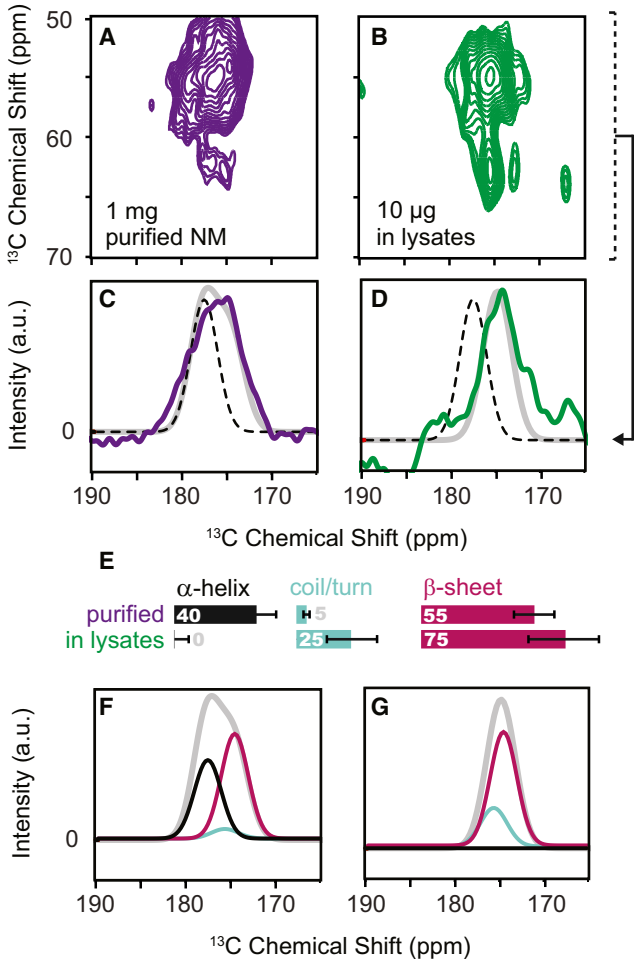


Figure 3. The Secondary Structure of NM Fibers in Cellular Lysates Differs from the Secondary Structure of In Vitro Templated NM

(A and B) Carbonyl carbon region of ^{13}C-^{13}C correlation spectra at 700 MHz using DNP MAS NMR of (A) cryoprotected purified NM fibers acquired in 6 hr and (B) cryoprotected NM fibers assembled in the presence of cellular lysates acquired in 1 week.

(C and D) Examination of the carbonyl carbon (C′) region of the spectra in projections of the Cα region (50–70 ppm indicated by dotted bracket) reveals the secondary structural composition of the protein backbone. The projection eliminates signals from non-backbone sites, such as the carbonyl moieties in the amino acid side chains like Asn and Gln. Dotted black lines indicate the expected chemical shift values for α-helical conformations of the protein backbone and highlight a large shift away from α-helical character for NM in lysates (D). The gray line represents the best-fitted solution to three Gaussian distributions describing the expected chemical shifts for the three possible secondary structural motifs: α helices (177.8 ± 1.5 ppm), random coils and turns (175.6 ± 1.5 ppm) and beta sheets (175.4 ± 1.55 ppm) (Wang and Jardetzky, 2002). Fits to a sum of these three Gaussian distributions gave standard estimates of error of 0.84 (C) and 0.93 (D). Residuals are plotted in Figure S3.

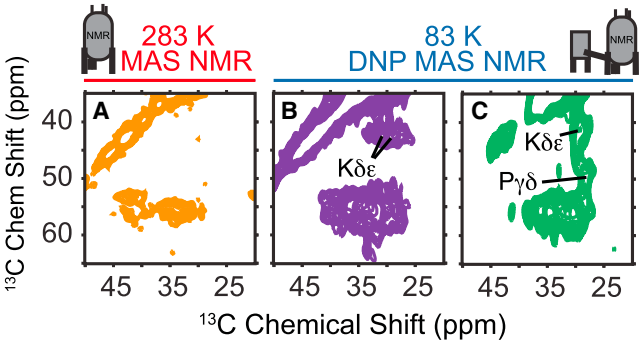
(E) Relative secondary structure contributions (in percent) as determined by intensity of each Gaussian distribution for the protein backbone of purified NM fibers (top) and NM fibers assembled in lysates (bottom). The error bars represent the standard error for the fitted intensity of each of the Gaussian distribution.

(F and G) The fitted intensities for α helices (black), random coils and turns (light blue) and beta sheets (magenta) are plotted with the fits (gray) from (C) and (D).

room temperature spectra (Figure S2). This establishes that DNP conditions did not compromise resolution gains at high magnetic fields, consistent with several other recent reports for cryogenic experiments on amyloid proteins (Debelouchina et al., 2010; Linden et al., 2011; Lopez del Amo et al., 2013)

Native Environments Structure Intrinsically Disordered Regions

Thus poised, we sought to determine the structural influences of the biological context on NM. The NMR chemical shift is a sensitive indicator of the secondary structure of the protein backbone. To investigate effects of lysates on NM secondary structure, we compared the backbone chemical shifts in the presence and absence of cellular lysates. To isolate signals from backbone C′-C_α sites, we projected the C_α region of the one-bond ^{13}C-^{13}C DARR correlation spectra into one dimension (Figure 3). We fit the carbonyl region of the projections to a sum of three Gaussians that described the chemical shift distributions



D

MSDSNQGNNQ QNYQQYSQNG NQQQGNNRYQ GYQAYNAQAQ
PAGGYYQNYQ GYSGYQQGGY QQYNPDAGYQ QQYNPQGGYQ
QYNPQGGYQQ QFNPQGGRGN YKNFNYNNNL QGYQAGFQPQ
SQGMSLNDFQ KQQKQAAPKP KKTLKLVSSS GIKLANATKK
VGTKPAESDK KEEEKSAETK EPTKEPTKVE EPVKKEEKPV
QTEEKTEEKS ELPKVEDLKI SESTHNTNNA NVTSADALIK
EQEEEVDDEV VND

Figure 4. Complex Biological Environments Restructure Intrinsically Disordered Protein Regions

(A–C) Side chains of NM fibers in cellular lysates have a different chemical environment than in vitro templated NM. Aliphatic region of (A) purified NM fibers at 283 K in protonated assembly buffer (B) purified NM fibers at 83 K in 60% d_8-glycerol and (C) NM fibers at 83 K in 60% d_8-glycerol templated into the amyloid form in the presence of cellular lysates. See also Figures S2 and S4.

(D) Amino acid sequence of NM with positions of lysines (magenta) and proline (cyan) highlighted. The N domain is black and the M domain is gray.

for α helices, random coil, and beta sheet conformations (Wang and Jardetzky, 2002) (Figure 3). At 283 K, NM fibers experience motion over a broad range of timescales (Frederick et al., 2014). The rigid regions of NM fibers at 283 K had a chemical shift distribution consistent with a mix of turns and sheets (Figure S3). Cryoprotected NM fibers at 83 K had a chemical shift distribution that was dramatically shifted toward α-helical values, consistent with sequence-based secondary structural predictions for the M domain (Chou and Fasman, 1974; Cuff et al., 1998; Kumar, 2013). This change is likely a result of secondary structural stabilization effects from the low experimental temperature and the cryoprotectant (Mehrnejad et al., 2011; Vagenende et al., 2009). In contrast, cryoprotected NM fibers that had been polymerized in cellular lysates had a chemical shift distribution that was dramatically shifted away from α-helical values and toward beta sheet values (Figure 3). Thus, the cellular context had a profound effect on protein secondary structure.

The NMR chemical shift is a sensitive indicator of chemical identity and structural conformation (Wang and Jardetzky, 2002). To determine which amino acid types undergo changes in their secondary structure in cellular contexts, we therefore compared the aliphatic region of the ^{13}C-^{13}C correlation spectra because this region reports on the chemically diverse amino acid side chains. The amyloid core of NM is largely composed of N, Q, and Y residues. In purified room temperature samples, the ^{13}C-^{13}C correlation spectra was consistent with an amyloid core containing N, Q, and Y residues in rigid beta sheet and turn conformations. Changes in the secondary structure at the α carbon for N, Q, and Y from a beta sheet or random coil conformation to an α-helical conformation result in an average change in chemical shift of ~4 ppm (Wang and Jardetzky, 2002). The average chemical shift values for this region of the spectra at 83 K for both purified NM and NM in cell lysates were the same as those for the room temperature sample, consistent with the amyloid character of NM being maintained at low temperatures and unperturbed by a biological context (Figures 4 and Figure S4). Thus, the secondary structural changes (Figure 3) were not derived from a structural rearrangement of the amyloid core.

We next compared the aliphatic region of the ^{13}C-^{13}C correlation spectra to determine if amino acid types found in the M domain of NM were affected the biological contexts. In purified room temperature spectra of NM, the ^{13}C-^{13}C correlation spectra was consistent with previous findings that established that the M domain is highly dynamic with random coil character (Frederick et al., 2014; Luckgei et al., 2013); side chain resonances for the amino acid types found only in the M domain were absent. At 83 K, cross-peaks for methyl-bearing amino acids such as threonine, valine, isoleucine and leucine found in the M domain were absent from both spectra due to temperature-dependent dynamically mediated relaxation of methyl-bearing amino acid side chains at this temperature (Bajaj et al., 2009; Beshah et al., 1987). However, at 83 K, lysine Cδ-Cε and proline Cγ-Cδ cross-peaks were present in the DNP MAS NMR spectra of both purified NM and of NM in cell lysates. Unlike the amyloid core residues, these amino acid types had very different chemical environments with differences in chemical shift of 5 ppm or greater depending on whether or not the fibers were templated in cellular lysates (Figure 4). Proline residues are present throughout the sequence of NM, while lysine residues are found only in the M domain (Figure 4D) localizing the regions experiencing large structural changes to the M domain. There are 25 lysine residues in the M domain of NM that contribute to the signal, all of which have different chemical environments and therefore different chemical shifts. The dramatic change in the shape of the lysine Cδ-Cε cross-peak indicates that a large proportion of the lysine side chains have a dramatically altered chemical environment in cellular lysates, indicating the majority of the M domain is involved. This establishes that the M domain, which contains chaperone-binding sites critical for faithful prion inheritance, makes many interactions with such components in vivo.

Multiple lines of evidence reveal that chaperone proteins directly interact with NM fibers. For example, the Hsp70 chaperone proteins Ssa1p and Ssa2p interact with NM aggregates (Allen et al., 2005), are among the top one hundred most highly expressed proteins (Ghaemmaghami et al., 2003) and the major components of amyloid aggregates isolated from yeast (Bagriantsev et al., 2008). In prion-containing cells, NM forms membrane-free cellular structures with specific cellular localizations (Tyedmers et al., 2010). Within these structures, NM amyloid fibers are deposited in highly ordered arrays of regularly spaced fibrils. These arrays consist of bundles of fibers organized by inter-fibrils structures that are thought to be an Hsp70 because cells lacking Hsp70 can no longer form ordered arrays (Saibil et al., 2012). This organization may be important for the faithful

inheritance of the prion by daughter cells or for mitigating the toxicity that is otherwise associated with protein aggregation. The direct observation of NM structure in its biological context indicates that these organizing protein-protein interactions are mediated through the M domain of the protein via the adoption of a beta sheet secondary structure by the majority of this otherwise intrinsically disordered region. This work suggests that disordered regions that are often observed in purified fibril samples may be intimately involved with cellular components to create a self-organization mechanism that coordinates fiber deposition.

DISCUSSION

Application of high-field DNP MAS NMR methodology to a challenging biological system allowed us to pursue a scientific question that was previously impossible due to limits in instrumental sensitivity. Without DNP, these experiments would not be possible. With DNP MAS NMR, we detected prion fibrils that had been assembled in a complex cellular environment containing all of the potential organizing protein components, such as chaperones, at their endogenous levels and stoichiometries. We established such fibers are structurally distinct from purified fibers in a region that is intrinsically disordered and highly dynamic in purified systems. The cellular environment structures an intrinsically disordered region. Sup35 is not unique; over a third of encoded proteins are predicted to be intrinsically disordered (Dunker et al., 2001). Indeed, intrinsically disordered protein regions have important roles in many biological processes, yet their structural characterization is notoriously difficult. Using DNP NMR, we can directly observe a protein of interest in its biological context. We found that the intrinsically disordered domain makes many direct interactions with cellular components. For NM, this suggests the M domain may be responsible for mediating interactions with the inter-fiber structures involved in prion fibril bundle organization visualized using in vivo cryotomography (Saibil et al., 2012).

Our results demonstrate not only that structural studies of proteins in their native contexts are possible, but also that the native context can and does have a dramatic influence on protein structure. We anticipate that our methodology will enable structural investigations of heterotypic quaternary interactions between a protein of interest and cellular constituents. The methods described in this work can be extended to further investigations of protein conformation in biologically relevant environments. For example, protein structures can be determined in cellular contexts that have been modified, either genetically by deletion or overexpression of a protein or by the addition of small molecule agonists. Moreover, because the protein of interest is prepared exogenously, the full suite of specific isotopic enrichment schemes can be employed (Jaipuria et al., 2012) or segmentally isotopically labeled proteins can be used to obtain atomic level structural insights for otherwise crowded spectra (Volkmann and Iwaï, 2010). These approaches will be particularly useful for structural investigations of protein folding and mis-folding in native and perturbed environments. There are a large number of protein folding diseases and work across many fields of study is continually uncovering genetic, physical and chemical modulators of their pathobiology. Our approach will allow direct observation of the structural consequence of such modulators. Thus, this work provides the framework to answer structural questions about the toxic and non-toxic conformations of disease-associated proteins in a way that is directly informed by genetic backgrounds and biological phenotypes. This will allow us to investigate how genetic backgrounds modify the energetic landscape of protein folding and will enable tight coupling of genotypes, phenotypes, and environments with specific structural arrangements.

EXPERIMENTAL PROCEDURES

Sample Preparation

Both untagged NM and C-terminally His^6 tagged NM were expressed and purified as described elsewhere (Serio et al., 1999). Uniformly labeled ^{13}C NM samples were prepared by growing BL21(DES)-Rosetta *Escherichia coli* in the presence of M9 media with 2 g L^{-1} D-glucose $^1H,^{13}C_6$ (Cambridge Isotope Labs). Purified, lysate-templated NM seeds for the purified fiber sample were prepared as described elsewhere (Frederick et al., 2014), using cell lysates from a strong [*PSI*$^+$] yeast strain. One milligram of purified denatured ^{13}C-labeled NM was diluted 120-fold out of 6 M GdHCL into 4 ml of lysis buffer (see below) containing 0.02 mg preformed fibers. The reaction was allowed to polymerize for 24 hr at 4°C and fibers were collected by ultracentrifugation at 430,000 × *g* for 1 hr. Bradford analysis revealed that removal of the supernatant decreased the total protein content of the sample by one-third. The pellet was resuspended in 60:30:10 (v/v/v) mixture of ^{13}C-depleted d_8-glycerol (99.9% ^{12}C):D_2O:H_2O (Rosay et al., 2010) containing 10 mM of the stable biradical TOTAPOL (Corzilius et al., 2014; Lange et al., 2012; Song et al., 2006).

Cell Lysate Samples for DNP

Phenotypically strong [*PSI*$^+$] yeast were grown in a 20 ml culture volume at 30°C to mid-log phase in YPD media made with protonated carbon sources and 100% D_2O. Because we use protonated carbon sources, the final deuteration level for the lysates is estimated to be 70% (Leiting et al., 1998). Cells maintained their [*PSI*$^+$] status in deuterated media (Figure 1A). Cells were collected by centrifugation (5 min, 4,000 × *g*) and washed once with water and once with D_2O. Pellets were suspended in 200 μl of lysis buffer (50 mM Tris-HCl pH 7.4, 200 mM NaCl, 2 mM TCEP, 5% d_8-^{13}C-depleted glycerol, 1 mM EDTA, 5 μg/ml of aprotinin and leupeptin and 100 μg/ml Roche protease inhibitor cocktail; lysis buffer was 80% [v/v] D_2O.) Cells were lysed by bead beating with 500 μm acid washed glass beads for 8 min at 4°C. After bead beating, the bottom of the Eppendorf tube was punctured with a 22G needle and the entire lysate mixture was transferred to a new tube. Purified denatured ^{13}C-labeled NM was diluted 150-fold out of 6 M GdHCl to a final concentration of 5 μM and the mixture was allowed to polymerize, quiescent, at 4°C for 24 hr. Unassembled NM was removed by centrifugation at 20,000 × *g* for 1 hr at 4°C and removal of the supernatant. The ~30 μl pellet was resuspended in 30 μl of 100% d_8-^{13}C-depleted glycerol containing 20 mM TOTAPOL and transferred to a 4 mm sapphire rotor. The final radical concentration was 10 mM (Corzilius et al., 2014) and the glycerol concentration was 60% (Rosay et al., 2010). The cell lysate sample for high field DNP was made analogously, except that yeast cells were grown in SD-CSM media made with D_2O and 2% (w/v) protonated ^{13}C-depleted glucose (99.9% ^{12}C, Cambridge Isotope Labs) as the carbon source. Uniform ^{13}C-labeled samples were grown using U-^{13}C glucose (99% Cambridge Isotope Labs) as the carbon source. The final sample volume was 20 μl and the sapphire rotor had a 3.2-mm diameter.

Immunohistochemistry

Cell lysate samples were made as described above, except NM-His^6 was substituted for NM. SDD-AGE was performed as described (Halfmann and Lindquist, 2008), and NM was visualized using an anti-His^6 antibody. Cell lysates were fractionated by SDS-PAGE, transferred to nitrocellulose and probed with both anti-His^6 and anti-Sup35 antibodies. For SDD-AGE analysis we prepared cellular lysates as described above and added 5 μM purified

denatured NM-His^6 to reactions containing cellular lysates from prion minus ([*psi*$^-$]) cultures, purified NM fibers prepared in isolation (2% seeding w/w), and cellular lysates from prion plus ([*PSI*$^+$]) cultures. For western blot analysis, lysate samples were denatured by incubation at 95°C for 10 min in the presence of 2% SDS before fractionation to denature amyloid aggregates. Secondary antibodies were coupled to horseradish peroxidase. Blots were visualized by a standard ECL analysis.

Spectroscopy

DNP MAS NMR experiments were performed on custom-designed home-built instruments, consisting of a 212 MHz (^{1}H, 5 T) (Becerra et al., 1993) and a 697 MHz (^{1}H, 16.4 T) (Barnes et al., 2012; Michaelis et al., 2014) NMR spectrometer (courtesy of Dr. David Ruben, Francis Bitter Magnet Laboratory, MIT) equipped with custom-built 140 and 460 GHz gyrotrons (Joye et al., 2006) (i.e., high power microwave devices generating up to 12 W), respectively. DNP MAS NMR spectra were recorded on home-built 4 mm (211 MHz) quadruple resonance (^{1}H, ^{13}C, ^{15}N, and e$^-$) or 3.2 mm (700 MHz) triple resonance (^{1}H, ^{13}C, and e$^-$) cryogenic probes equipped with Kel-F stators (Revolution NMR). Microwaves were guided to the sample via circular overmoded waveguide in which the inner surface has been corrugated to reduce mode conversion and ohmic losses. Sample temperatures were maintained below 85 K, with spinning frequencies of $\omega_r/2\pi$ = 4.3 – 10 kHz.

^{13}C{^{1}H} cross polarization (Pines et al., 1973) spectra were acquired with a contact time of 1.5 ms. Recycle delays were chosen as T_B (polarization buildup time constant) × 1.26 (Figure S1), yielding optimum sensitivity per unit of time. The recycle delays were 4.6 s and 8 s for 211 MHz and 700 MHz, respectively. A series of ^{13}C-^{13}C DARR spectra were recorded using either a mixing period of 6 or 15 ms, 64–512 co-added transients and, between 60 and 100 t_2 increments. All data were acquired using high-power TPPM ^{1}H decoupling ($\gamma B_1 > 83$ kHz). Enhancements at 211 MHz are reported in Figure 2 and those at 700 MHz were estimated at −8 to −10. DNP enhancements at both fields were ~80% of the maximal enhancements recorded on a standard sample of proline. Experimental data were processed using RNMR (1D) or NMRpipe (Delaglio et al., 1995) (2D) and analyzed using Sparky (Goddard and Kneller, 2006). ^{13}C NMR data were referenced to adamantane (Morcombe and Zilm, 2003) (40.49 ppm at room temperature), and KBr was used to set the magic angle.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.024.

AUTHOR CONTRIBUTIONS

Conceptualization, K.K.F.; Methodology, K.K.F.; Investigation, K.K.F., A.J., V.K.M., B.C., and T.-C.O. Writing-Original Draft, K.K.F.; Writing-Review and Editing, K.K.F., V.K.M., and S.L.; Funding, S.L. and R.G.G.; Supervision, K.K.F., S.L., and R.G.G.

ACKNOWLEDGMENTS

We thank the members of the S.L. and R.G.G. groups for valuable discussions and comments during the course of this research. S.L. is an investigator of the Howard Hughes Medical Institute. K.K.F. was supported by the Life Science Research Foundation as an HHMI fellow. V.K.M. is grateful to the Natural Sciences and Engineering Research Council of Canada and the Government of Canada for a Banting postdoctoral fellowship. B.C. was supported by the Deutsche Forschungsgemeinschaft (research fellowship CO 802/1-1). This work was funded by grants from the G. Harold and Leila Y. Mathers Foundation (S.L.) and by NIH grants GM-025874 to S.L. and EB-003151, EB-002804, and EB-002026 to R.G.G.

Received: May 4, 2015
Revised: July 3, 2015
Accepted: August 26, 2015
Published: October 8, 2015

REFERENCES

Abragam, A. (1983). The Principles of Nuclear Magnetism (Clarendon Press).

Akbey, Ü., Franks, W.T., Linden, A., Lange, S., Griffin, R.G., van Rossum, B.-J., and Oschkinat, H. (2010). Dynamic nuclear polarization of deuterated proteins. Angew. Chem. Int. Ed. Engl. *49*, 7803–7806.

Akbey, Ü., Franks, W.T., Linden, A., Orwick-Rydmark, M., Lange, S., and Oschkinat, H. (2013). Dynamic nuclear polarization enhanced NMR in the solid-state. Top. Curr. Chem. *338*, 181–228.

Allen, K.D., Wegrzyn, R.D., Chernova, T.A., Müller, S., Newnam, G.P., Winslett, P.A., Wittich, K.B., Wilkinson, K.D., and Chernoff, Y.O. (2005). Hsp70 chaperones as modulators of prion life cycle: novel effects of Ssa and Ssb on the Saccharomyces cerevisiae prion [PSI+]. Genetics *169*, 1227–1242.

Bagriantsev, S.N., Kushnirov, V.V., and Liebman, S.W. (2006). Analysis of amyloid aggregates using agarose gel electrophoresis. Methods Enzymol. *412*, 33–48.

Bagriantsev, S.N., Gracheva, E.O., Richmond, J.E., and Liebman, S.W. (2008). Variant-specific [PSI+] infection is transmitted by Sup35 polymers within [PSI+] aggregates with heterogeneous protein composition. Mol. Biol. Cell *19*, 2433–2443.

Bajaj, V.S., van der Wel, P.C.A., and Griffin, R.G. (2009). Observation of a low-temperature, dynamically driven structural transition in a polypeptide by solid-state NMR spectroscopy. J. Am. Chem. Soc. *131*, 118–128.

Banci, L., Barbieri, L., Bertini, I., Luchinat, E., Secci, E., Zhao, Y., and Aricescu, A.R. (2013). Atomic-resolution monitoring of protein maturation in live human cells by NMR. Nat. Chem. Biol. *9*, 297–299.

Barnes, A.B., Markhasin, E., Daviso, E., Michaelis, V.K., Nanni, E.A., Jawla, S.K., Mena, E.L., DeRocher, R., Thakkar, A., Woskov, P.P., et al. (2012). Dynamic nuclear polarization at 700 MHz/460 GHz. J. Magn. Reson. *224*, 1–7.

Becerra, L.R., Gerfen, G.J., Temkin, R.J., Singel, D.J., and Griffin, R.G. (1993). Dynamic nuclear polarization with a cyclotron resonance mazer at 5 T. Phys. Rev. Lett. *71*, 3561–3564.

Beshah, K., Olejniczak, E.T., and Griffin, R.G. (1987). Deuterium NMR study of methyl group dynamics in L-alanine. J. Chem. Phys. *86*, 4730.

Chien, P., Weissman, J.S., and DePace, A.H. (2004). Emerging principles of conformation-based prion inheritance. Annu. Rev. Biochem. *73*, 617–656.

Chou, P.Y., and Fasman, G.D. (1974). Prediction of protein conformation. Biochemistry *13*, 222–245.

Corzilius, B., Andreas, L.B., Smith, A.A., Ni, Q.Z., and Griffin, R.G. (2014). Paramagnet induced signal quenching in MAS-DNP experiments in frozen homogeneous solutions. J. Magn. Reson. *240*, 113–123.

Cox, B.S. (1965). Ψ, A cytoplasmic suppressor of super-suppressor in yeast. Heredity *20*, 505–521.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., and Barton, G.J. (1998). JPred: a consensus secondary structure prediction server. Bioinformatics *14*, 892–893.

Debelouchina, G.T., Bayro, M.J., van der Wel, P.C.A., Caporini, M.A., Barnes, A.B., Rosay, M., Maas, W.E., and Griffin, R.G. (2010). Dynamic nuclear polarization-enhanced solid-state NMR spectroscopy of GNNQQNY nanocrystals and amyloid fibrils. Phys. Chem. Chem. Phys. *12*, 5911–5919.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293.

Dobson, C.M. (2001). The structural basis of protein folding and its links with human disease. Philos. Trans. R. Soc. Lond. B Biol. Sci. *356*, 133–145.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. (2001). Intrinsically disordered protein. J. Mol. Graph. Model. *19*, 26–59.

Frederick, K.K., Debelouchina, G.T., Kayatekin, C., Dorminy, T., Jacavone, A.C., Griffin, R.G., and Lindquist, S. (2014). Distinct prion strains are defined by amyloid core structure and chaperone binding site dynamics. Chem. Biol. *21*, 295–305.

Freedberg, D.I., and Selenko, P. (2014). Live cell NMR. Annu. Rev. Biophys. *43*, 171–192.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Goddard, T.D., and Kneller, D.G. (2006). Sparky (University of California).

Guo, J.L., Covell, D.J., Daniels, J.P., Iba, M., Stieber, A., Zhang, B., Riddle, D.M., Kwong, L.K., Xu, Y., Trojanowski, J.Q., and Lee, V.M. (2013). Distinct α-synuclein strains differentially promote tau inclusions in neurons. Cell *154*, 103–117.

Halfmann, R., and Lindquist, S. (2008). Screening for amyloid aggregation by semi-denaturing detergent-agarose gel electrophoresis. J. Vis. Exp. *17*, 838.

Heise, H., Hoyer, W., Becker, S., Andronesi, O.C., Riedel, D., and Baldus, M. (2005). Molecular-level secondary structure, polymorphism, and dynamics of full-length alpha-synuclein fibrils studied by solid-state NMR. Proc. Natl. Acad. Sci. USA *102*, 15871–15876.

Helmus, J.J., Surewicz, K., Nadaud, P.S., Surewicz, W.K., and Jaroniec, C.P. (2008). Molecular conformation and dynamics of the Y145Stop variant of human prion protein in amyloid fibrils. Proc. Natl. Acad. Sci. USA *105*, 6284–6289.

Helsen, C.W., and Glover, J.R. (2012). Insight into molecular basis of curing of [PSI+] prion by overexpression of 104-kDa heat shock protein (Hsp104). J. Biol. Chem. *287*, 542–556.

Inomata, K., Ohno, A., Tochio, H., Isogai, S., Tenno, T., Nakase, I., Takeuchi, T., Futaki, S., Ito, Y., Hiroaki, H., and Shirakawa, M. (2009). High-resolution multi-dimensional NMR spectroscopy of proteins in human cells. Nature *458*, 106–109.

Jacso, T., Franks, W.T., Rose, H., Fink, U., Broecker, J., Keller, S., Oschkinat, H., and Reif, B. (2012). Characterization of membrane proteins in isolated native cellular membranes by dynamic nuclear polarization solid-state NMR spectroscopy without purification and reconstitution. Angew. Chem. Int. Ed. Engl. *51*, 432–435.

Jaipuria, G., Krishnarjuna, B., Mondal, S., Dubey, A., and Atreya, H.S. (2012). Amino acid selective labeling and unlabeling for protein resonance assignments. Adv. Exp. Med. Biol. *992*, 95–118.

Joye, C.D., Griffin, R.G., Hornstein, M.K., Hu, K.-N., Kreischer, K.E., Rosay, M., Shapiro, M.A., Sirigiri, J.R., Temkin, R.J., and Woskov, P.P. (2006). Operational characteristics of a 14-W 140-GHz gyrotron for dynamic nuclear polarization. IEEE Trans Plasma Sci IEEE Nucl Plasma Sci Soc *34*, 518–523.

Jucker, M., and Walker, L.C. (2013). Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. Nature *501*, 45–51.

Kiktev, D.A., Patterson, J.C., Müller, S., Bariar, B., Pan, T., and Chernoff, Y.O. (2012). Regulation of chaperone effects on a yeast prion by cochaperone Sgt2. Mol. Cell. Biol. *32*, 4960–4970.

Kodali, R., Williams, A.D., Chemuru, S., and Wetzel, R. (2010). Aβ(1–40) Forms five distinct amyloid structures whose β-sheet contents and fibril stabilities are correlated. J. Mol. Biol. *401*, 503–517.

Krishnan, R., and Lindquist, S.L. (2005). Structural insights into a yeast prion illuminate nucleation and strain diversity. Nature *435*, 765–772.

Kumar, T.A. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction Server. Wide Spectrum *1*, 15–19.

Lange, S., Linden, A.H., Akbey, Ü., Franks, W.T., Loening, N.M., van Rossum, B.-J., and Oschkinat, H. (2012). The effect of biradical concentration on the performance of DNP-MAS-NMR. J. Magn. Reson. *216*, 209–212.

Leiting, B., Marsilio, F., and O'Connell, J.F. (1998). Predictable deuteration of recombinant proteins expressed in Escherichia coli. Anal. Biochem. *265*, 351–355.

Linden, A.H., Franks, W.T., Akbey, Ü., Lange, S., van Rossum, B.-J., and Oschkinat, H. (2011). Cryogenic temperature effects and resolution upon slow cooling of protein preparations in solid state NMR. J. Biomol. NMR *51*, 283–292.

Liu, J.-J., Sondheimer, N., and Lindquist, S.L. (2002). Changes in the middle region of Sup35 profoundly alter the nature of epigenetic inheritance for the yeast prion [PSI+]. Proc. Natl. Acad. Sci. USA *99* (*Suppl 4*), 16446–16453.

Lopez del Amo, J.-M., Schneider, D., Loquet, A., Lange, A., and Reif, B. (2013). Cryogenic solid state NMR studies of fibrils of the Alzheimer's disease amyloid-β peptide: perspectives for DNP. J. Biomol. NMR *56*, 359–363.

Luckgei, N., Schütz, A.K., Bousset, L., Habenstein, B., Sourigues, Y., Gardiennet, C., Meier, B.H., Melki, R., and Böckmann, A. (2013). The conformation of the prion domain of Sup35p in isolation and in the full-length protein. Angew. Chem. Int. Ed. Engl. *52*, 12741–12744.

Masison, D.C., Kirkland, P.A., and Sharma, D. (2009). Influence of Hsp70s and their regulators on yeast prion propagation. Prion *3*, 65–73.

Mehrnejad, F., Ghahremanpour, M.M., Khadem-Maaref, M., and Doustdar, F. (2011). Effects of osmolytes on the helical conformation of model peptide: molecular dynamics simulation. J. Chem. Phys. *134*, 035104.

Michaelis, V.K., Ong, T.-C., Kiesewetter, M.K., Frantz, D.K., Walish, J.J., Ravera, E., Luchinat, C., Swager, T.M., and Griffin, R.G. (2014). Topical Developments in High-Field Dynamic Nuclear Polarization. Isr. J. Chem. *54*, 207–221.

Morcombe, C.R., and Zilm, K.W. (2003). Chemical shift referencing in MAS solid state NMR. J. Magn. Reson. *162*, 479–486.

Nekooki-Machida, Y., Kurosawa, M., Nukina, N., Ito, K., Oda, T., and Tanaka, M. (2009). Distinct conformations of in vitro and in vivo amyloids of huntingtin-exon1 show different cytotoxicity. Proc. Natl. Acad. Sci. USA *106*, 9679–9684.

Ni, Q.Z., Daviso, E., Can, T.V., Markhasin, E., Jawla, S.K., Swager, T.M., Temkin, R.J., Herzfeld, J., and Griffin, R.G. (2013). High frequency dynamic nuclear polarization. Acc. Chem. Res. *46*, 1933–1941.

Petkova, A.T., Leapman, R.D., Guo, Z., Yau, W.-M., Mattson, M.P., and Tycko, R. (2005). Self-propagating, molecular-level polymorphism in Alzheimer's beta-amyloid fibrils. Science *307*, 262–265.

Pines, A., Gibby, M.G., and Waugh, J.S. (1973). Proton-enhanced NMR of dilute spins in solids. J. Chem. Phys. *59*, 569.

Polymenidou, M., and Cleveland, D.W. (2012). Prion-like spread of protein aggregates in neurodegeneration. J. Exp. Med. *209*, 889–893.

Prusiner, S.B., Scott, M.R., DeArmond, S.J., and Cohen, F.E. (1998). Prion protein biology. Cell *93*, 337–348.

Reckel, S., Lopez, J.J., Löhr, F., Glaubitz, C., and Dötsch, V. (2012). In-cell solid-state NMR as a tool to study proteins in large complexes. ChemBioChem *13*, 534–537.

Renault, M., Pawsey, S., Bos, M.P., Koers, E.J., Nand, D., Tommassen-van Boxtel, R., Rosay, M., Tommassen, J., Maas, W.E., and Baldus, M. (2012). Solid-state NMR spectroscopy on cellular preparations enhanced by dynamic nuclear polarization. Angew. Chem. Int. Ed. Engl. *51*, 2998–3001.

Rosay, M., Tometich, L., Pawsey, S., Bader, R., Schauwecker, R., Blank, M., Borchard, P.M., Cauffman, S.R., Felch, K.L., Weber, R.T., Temkin, R.J., Griffin, R.G., and Maas, W.E. (2010). Solid-state dynamic nuclear polarization at 263 GHz: spectrometer design and experimental results. Phys. Chem. Chem. Phys. *12*, 5850–5860.

Saibil, H.R., Seybert, A., Habermann, A., Winkler, J., Eltsov, M., Perkovic, M., Castaño-Diez, D., Scheffer, M.P., Haselmann, U., Chlanda, P., et al. (2012). Heritable yeast prions have a highly organized three-dimensional architecture with interfiber structures. Proc. Natl. Acad. Sci. USA *109*, 14906–14911.

Sakakibara, D., Sasaki, A., Ikeya, T., Hamatsu, J., Hanashima, T., Mishima, M., Yoshimasu, M., Hayashi, N., Mikawa, T., Wälchli, M., et al. (2009). Protein structure determination in living cells by in-cell NMR spectroscopy. Nature *458*, 102–105.

Selenko, P., Serber, Z., Gadea, B., Ruderman, J., and Wagner, G. (2006). Quantitative NMR analysis of the protein G B1 domain in Xenopus laevis egg extracts and intact oocytes. Proc. Natl. Acad. Sci. USA *103*, 11904–11909.

Serio, T.R., Cashikar, A.G., Moslehi, J.J., Kowal, A.S., and Lindquist, S.L. (1999). Yeast prion [psi +] and its determinant, Sup35p. Methods Enzymol. *309*, 649–673.

Slichter, C.P. (1990). Principles of Magnetic Resonance (Springer Science and Business Media).

Song, C., Hu, K.-N., Joo, C.-G., Swager, T.M., and Griffin, R.G. (2006). TOTAPOL: a biradical polarizing agent for dynamic nuclear polarization experiments in aqueous media. J. Am. Chem. Soc. *128*, 11385–11390.

Szeverenyi, N.M., Sullivan, M.J., and Maciel, G.E. (1982). Observation of spin exchange by two-dimensional fourier transform 13C cross polarization-magic-angle spinning. J. Magn. Reson. (1969) *47*, 462–475.

Takahashi, H., Fernández-de-Alba, C., Lee, D., Maurel, V., Gambarelli, S., Bardet, M., Hediger, S., Barra, A.-L., and De Paëpe, G. (2014). Optimization of an absolute sensitivity in a glassy matrix during DNP-enhanced multidimensional solid-state NMR experiments. J. Magn. Reson. *239*, 91–99.

Takegoshi, K., Nakamura, S., and Terao, T. (2001). ^{13}C-^{1}H dipolar-assisted rotational resonance in magic-angle spinning NMR. Chem. Phys. Lett. *344*, 631–637.

Toyama, B.H., Kelly, M.J.S., Gross, J.D., and Weissman, J.S. (2007). The structural basis of yeast prion strain variants. Nature *449*, 233–237.

Tuite, M.F., Marchante, R., and Kushnirov, V. (2011). Fungal prions: structure, function and propagation. Top. Curr. Chem. *305*, 257–298.

Tyedmers, J., Treusch, S., Dong, J., McCaffery, J.M., Bevis, B., and Lindquist, S. (2010). Prion induction involves an ancient system for the sequestration of aggregated proteins and heritable changes in prion fragmentation. Proc. Natl. Acad. Sci. USA *107*, 8633–8638.

Uversky, V.N. (2013). A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. *22*, 693–724.

Vagenende, V., Yap, M.G.S., and Trout, B.L. (2009). Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. Biochemistry *48*, 11084–11096.

Vaiphei, S.T., Tang, Y., Montelione, G.T., and Inouye, M. (2011). The use of the condensed single protein production system for isotope-labeled outer membrane proteins, OmpA and OmpX in E. coli. Mol. Biotechnol. *47*, 205–210.

Volkmann, G., and Iwaï, H. (2010). Protein trans-splicing and its use in structural biology: opportunities and limitations. Mol. Biosyst. *6*, 2110–2121.

Wang, Y., and Jardetzky, O. (2002). Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci. *11*, 852–861.

Wang, T., Park, Y.B., Caporini, M.A., Rosay, M., Zhong, L., Cosgrove, D.J., and Hong, M. (2013). Sensitivity-enhanced solid-state NMR detection of expansin's target in plant cell walls. Proc. Natl. Acad. Sci. USA *110*, 16444–16449.

Wasmer, C., Schütz, A., Loquet, A., Buhtz, C., Greenwald, J., Riek, R., Böckmann, A., and Meier, B.H. (2009). The molecular organization of the fungal prion HET-s in its amyloid form. J. Mol. Biol. *394*, 119–127.

Watts, J.C., Giles, K., Oehler, A., Middleton, L., Dexter, D.T., Gentleman, S.M., DeArmond, S.J., and Prusiner, S.B. (2013). Transmission of multiple system atrophy prions to transgenic mice. Proc. Natl. Acad. Sci. USA *110*, 19555–19560.

Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins incellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29.

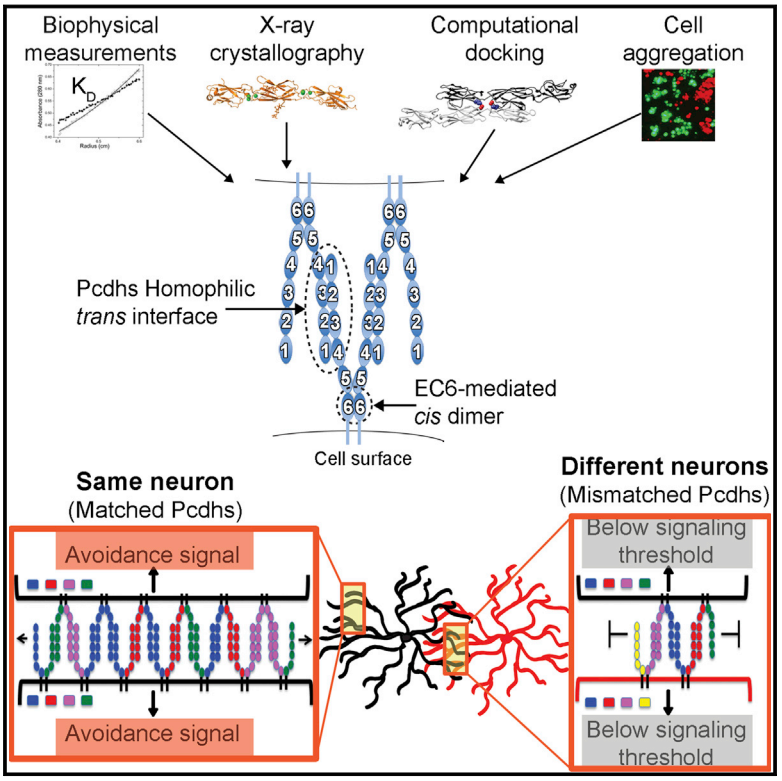
Yamamoto, K., Caporini, M.A., Im, S.-C., Waskell, L., and Ramamoorthy, A. (2015). Cellular solid-state NMR investigation of a membrane protein using dynamic nuclear polarization. Biochim. Biophys. Acta *1848*, 342–349.

Article

Cell

Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions

Graphical Abstract



Authors

Rotem Rubinstein, Chan Aye Thu, Kerry Marie Goodman, ..., Tom Maniatis, Lawrence Shapiro, Barry Honig

Correspondence

lss8@columbia.edu (L.S.), bh6@columbia.edu (B.H.)

In Brief

Protocadherin isoforms mediate neuronal self-recognition through a zipper-like association mechanism that allows recognition of isoform mismatch and chain-termination of the interactions.

Highlights

- Crystal structures of EC1–EC3 regions of Pcdh-α, -β, and -γ isoforms are determined
- Pcdh homophilic specificity is mediated by a canonical EC1–EC4 domain interface
- Pcdhs dimerize in *cis* through the EC6 domain independent of their *trans* interactions
- Isoform-mismatch chain-termination mechanism to distinguish self from non-self

Accession Numbers

4ZPO
4ZPQ
4ZPP
4ZPN
4ZPM
4ZPL
4ZPS

Rubinstein et al., 2015, Cell *163*, 629–642
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.026

Article

Cell

Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions

Rotem Rubinstein,[1,2,7] Chan Aye Thu,[1,7] Kerry Marie Goodman,[1,7] Holly Noelle Wolcott,[1,7] Fabiana Bahna,[1,5] Seetha Mannepalli,[1] Goran Ahlsen,[2,5] Maxime Chevee,[1] Adnan Halim,[6] Henrik Clausen,[6] Tom Maniatis,[1] Lawrence Shapiro,[1,2,*] and Barry Honig[1,2,3,4,5,*]

[1]Department of Biochemistry and Molecular Biophysics
[2]Department of Systems Biology
[3]Department of Medicine
[4]Zuckerman Mind Brain Behavior Institute
[5]Howard Hughes Medical Institute
Columbia University, New York, NY 10032, USA
[6]Copenhagen Center for Glycomics, Departments of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark
[7]Co-first author
*Correspondence: lss8@columbia.edu (L.S.), bh6@columbia.edu (B.H.)
http://dx.doi.org/10.1016/j.cell.2015.09.026

SUMMARY

Self-avoidance, a process preventing interactions of axons and dendrites from the same neuron during development, is mediated in vertebrates through the stochastic single-neuron expression of clustered protocadherin protein isoforms. Extracellular cadherin (EC) domains mediate isoform-specific homophilic binding between cells, conferring cell recognition through a poorly understood mechanism. Here, we report crystal structures for the EC1–EC3 domain regions from four protocadherin isoforms representing the α, β, and γ subfamilies. All are rod shaped and monomeric in solution. Biophysical measurements, cell aggregation assays, and computational docking reveal that *trans* binding between cells depends on the EC1–EC4 domains, which interact in an antiparallel orientation. We also show that the EC6 domains are required for the formation of *cis*-dimers. Overall, our results are consistent with a model in which protocadherin *cis*-dimers engage in a head-to-tail interaction between EC1–EC4 domains from apposed cell surfaces, possibly forming a zipper-like protein assembly, and thus providing a size-dependent self-recognition mechanism.

INTRODUCTION

The human brain is composed of ~10 billion neurons, each of which can connect with up to thousands of others. Neuronal self-avoidance is a process in which dendrites and axons originating from the same neuron repel one another but can freely interact with neurites from other neurons. The combined properties of self-recognition and non-self-discrimination require that contacting neurons display diverse cell-surface identities that allow for discrimination between self and non-self (Hattori et al., 2009; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010).

In *Drosophila* and other invertebrates, self-avoidance is mediated by Dscam1 proteins—immunoglobulin superfamily members produced by alternative splicing of the *DSCAM1* pre-mRNA. This cell-autonomous and stochastic alternative splicing can theoretically produce up to 19,008 Dscam1 isoforms with distinct ectodomains, each of which have highly specific homophilic *trans* binding specificity (Hattori et al., 2008; Miura et al., 2013; Schmucker et al., 2000; Wojtowicz et al., 2007). Distinct cell-surface identities are generated in *Drosophila* by the stochastic expression of a small set of Dscam1 isoforms in each neuron (Miura et al., 2013). Homophilic interactions between identical sets of protein isoforms on the surface of neurites from the same neuron result in repulsion and neurite self-avoidance (Hattori et al., 2008). The expression of even a single Dscam1 isoform is sufficient for self-avoidance of neurites from the same neuron (Hughes et al., 2007; Matthews et al., 2007; Soba et al., 2007). However, robust non-self-discrimination, which allows processes from different neurons to freely interact, requires thousands of distinct Dscam1 isoforms (Hattori et al., 2009).

Recent studies suggest that, in vertebrate nervous systems, neuronal self-avoidance functionality is provided, at least in part, by the clustered protocadherins (Pcdhs) (Chen and Maniatis, 2013; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010). Mammalian Pcdhs are encoded in a contiguous genomic locus composed of three adjacent gene clusters (*Pcdh* α, β, and γ), each of which contains close to 60 "variable" exons (58 in mice, Figure 1A) (Wu and Maniatis, 1999). Only a few variable exons are stochastically chosen for expression in each cell by a mechanism involving alternative promoter choice (Ribich et al., 2006; Tasic et al., 2002). Each variable exon encodes an entire Pcdh ectodomain region consisting of six tandem extracellular cadherin (EC) domains, a single transmembrane region, and a short cytoplasmic region. In the α and γ gene clusters, a "constant" C-terminal cytoplasmic region encoding an intracellular domain (ICD) is joined to the variable ectodomain

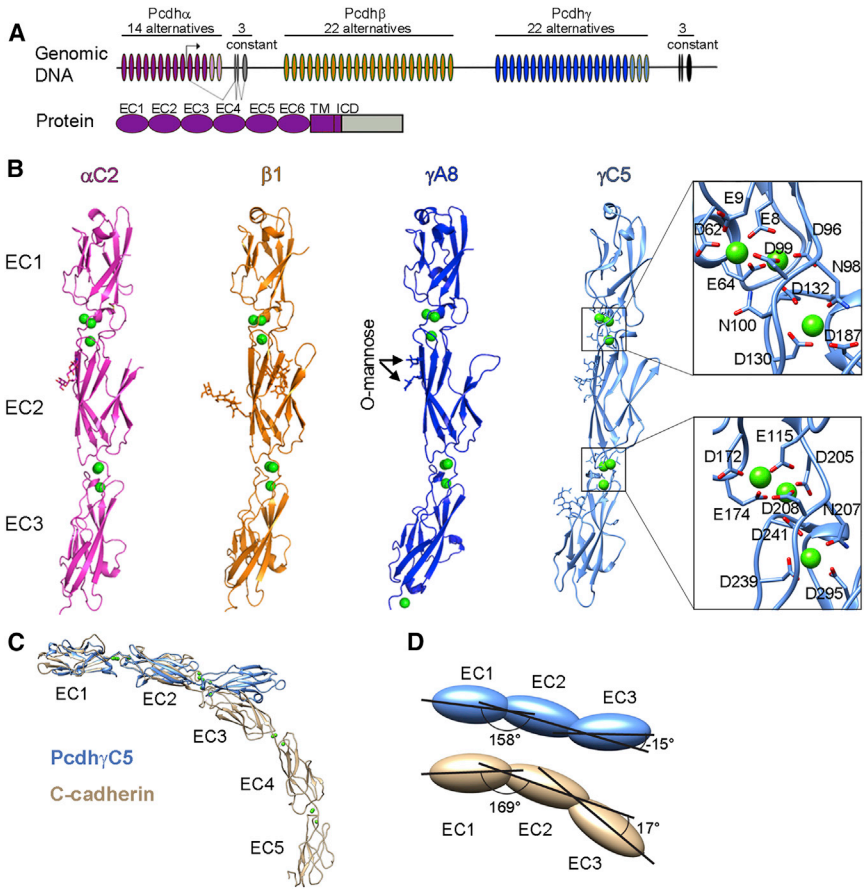


Figure 1. Crystal Structures of Four Pcdh EC1–EC3 Isoforms

(A) The Pcdh genomic locus contains three adjacent clusters of variable exons. Each exon encodes an entire ectodomain comprising six EC domains, a transmembrane (TM) domain, and a short cytoplasmic region. Alpha and gamma clusters also contain three constant exons that encode a cluster-specific ICD, which are joined by pre-mRNA splicing for alpha and gamma clusters. C-type Pcdh exons are shown in pink and light blue for the alpha and gamma clusters, respectively.

(B) Crystal structures of EC1–EC3 regions from PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 shown in ribbon representation. Ca^{2+} ions are drawn as green spheres. N-glycans and conserved O-mannose residues are drawn as sticks. The inter-domain calcium binding sites are arranged similarly to those observed in classical cadherins (expanded view). See also Figure S1 and Table S1.

(C) Comparison of the PcdhγC5 and type I classical C-cadherin structures. The overall architecture of classical cadherin ectodomains has a curved shape with an approximate 90° angle between EC1 and EC5 (Boggon et al., 2002). In contrast, the architecture of Pcdh EC1–EC3 domain regions is characterized by an extended zigzagged conformation.

(D) EC2–EC3 angles distinct from classical cadherins account for the extended zigzagged conformation of the Pcdh structures. EC1–EC3 domains are drawn as blue (PcdhγC5) and yellow (C-cadherin) ovals. Angles shown are between principal axes of inertia for adjacent domains.

exon by pre-mRNA splicing. The β cluster does not contain such a constant region, and therefore, β-Pcdhs are lacking an ICD. The α and γ gene clusters also encode a small set of "C-type" Pcdhs, which are divergent from other members of their respective clusters and appear to have distinct functions (Figure 1A) (Chen et al., 2012). Deletion of the *Pcdhγ* gene cluster in mice leads to the disruption of self-avoidance in retinal starburst amacrine cells and Purkinje cells with phenotypes similar to those described for *Dscam1* deletion mutants in *Drosophila* (Lefebvre et al., 2012).

Like invertebrate Dscam proteins, Pcdh isoforms engage in isoform-specific *trans* homophilic interactions (Schreiner and Weiner, 2010; Thu et al., 2014). It is remarkable that Pcdhs, with only 58 isoforms, can mediate neural self-recognition and non-self-discrimination similar to Dscams, which have up to tens of thousands of distinct extracellular isoforms. Central to this capability is the observation that a single mismatched Pcdh isoform can interfere with recognition between cells that express an otherwise matching set of Pcdhs (Thu et al., 2014). Understanding the mechanism underlying this "interference" phenomenon is crucial, as it is likely to explain how only 58 Pcdh isoforms can provide sufficient functional diversity to enable self-recognition and non-self-discrimination in the nervous system comparable to the much more diverse *Drosophila Dscam* gene.

Here, we report crystal structures of Pcdh extracellular protein fragments comprising the previously mapped Pcdh specificity-determining EC1–EC3 domains for PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 isoforms, thus providing examples from all three *Pcdh* gene clusters. Guided by these structures, we used two orthogonal mutagenesis approaches—surface-saturating arginine mutagenesis and bioinformatics-derived predictions—to map the isoform specificity-determining regions at the amino acid level using cell aggregation and biophysical experiments as readouts. The two approaches yielded consistent results, revealing an essential role for EC1 through EC4 in *trans* homophilic interactions and for EC6 in *cis* interactions. On the basis of these findings, we propose a model for Pcdh-mediated cell-cell recognition that is consistent with the remarkable ability of these cell-surface proteins to provide diverse single-cell identities to vertebrate neurons.

RESULTS

Structures of Pcdh EC1–EC3 Region Fragments from α, β, and γ Sub-families

We determined crystal structures of proteins composed of the three N-terminal EC domains of mouse PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 to a resolution of 2.4 Å, 3.3 Å, 2.9 Å, and 2.9 Å, respectively (Figure 1B and Table S1). We focused

Table 1. Analytical Ultracentrifugation Analysis of Clustered-Protocadherins Homo-oligomerization

Protein	Oligomeric State	K_D Oligomerization (μM)
$\alpha7_{EC1–EC3}$	monomer	NA
$\alpha C2_{EC1–EC3}$	non-specific dimer	242 ± 0.1[a]
$\beta1_{EC1–EC3}$	monomer	NA
$\gamma A8_{EC1–EC3}$	disulfide-linked dimer	NA
$\gamma C5_{EC1–EC3}$	monomer	NA
$\gamma C5_{EC1–EC3}$ extended N-term	monomer	NA
$\alpha7_{EC1–EC5}$	dimer	2.9 ± 0.5
$\alpha C2_{EC1–EC4}$	dimer	20 ± 1.2
$\alpha C2_{EC1–EC5}$	dimer	5.9 ± 0.8
$\gamma C5_{EC1–EC5}$	dimer	100 ± 4.3
$\gamma B6_{EC1–EC4}$	dimer	29 ± 4.9
$\gamma A8_{EC1–EC4}$	dimer	30 ± 1.5
$\gamma C5_{EC2–EC6}$	dimer	18 ± 0.2
$\gamma A8_{EC2–EC6}$	dimer	23 ± 8.1
$\alpha C2_{EC2–EC6}$	dimer	8.9 ± 0.3
$\alpha C2_{EC1–EC6}$	tetramer	0.1[b]
$\gamma C5_{EC1–EC6}$	tetramer	7.6[b]
$\gamma B6_{EC1–EC6}$	tetramer	0.2
$\gamma A8_{EC1–EC4}$ I116R	monomer	NA
$\gamma C5_{EC2–EC6}$ S116R	dimer	14
$\alpha C2_{EC1–EC6}$ S118R	tetramer	1.8[b]
$\gamma C5_{EC1–EC6}$ S116R	dimer	5.7

[a] n = 2; Isodesmic Ki = 359 μM; Ki/K_D = 1.48.
[b] K_D of a tetramer was obtained by locking the *cis*-interaction K_D as obtained from EC2–EC6 deletion constructs.

on protein fragments containing EC1–EC3, since the results of earlier cell aggregation experiments indicated that Pcdh isoform-specific recognition was mediated via the EC2–EC3 domains and that the EC1 domain is required for *trans* binding (Schreiner and Weiner, 2010).

The four structures show high overall similarity (Figures 1B and S1A). Each structure consists of three EC domains, each with the two-layer β sheet fold observed in classical cadherins. Successive domains are connected by calcium-binding linkers, each of which coordinate three Ca^{2+} ions utilizing side chains in the same conserved motifs (Figure 1B). These motifs are also conserved within type I and type II classical cadherins with the exception of the EE motif (bottom of EC1 domain, Figure 1B), which is present only in type-II cadherins. In contrast with previous conclusions (Schreiner and Weiner, 2010) but consistent with the presence of Ca^{2+} at the inter-domain linkers and in common with classical cadherins, we have found that cell aggregation of Pcdhs is Ca^{2+} dependent (Figure S1B). Despite these similarities to classical cadherins, the Pcdh isoform structures are distinctive in several aspects. Most notably, the overall arrangement of the three EC domains in each structure is much straighter than the curved classical cadherin architecture (Figure 1C). This "straight-rod" architecture arises from an extended zigzagged conformation: an arrangement that is generated primarily by a very different EC2–EC3 angle than classical cadherins (> 31° difference, Figure 1D).

In addition, mass spectrometry analyses showed that all four isoforms contain two sites of O-mannosylation at residues 194 and 196 (PcdhγC5 sequence numbering; Figures 1B, S1G, and S1H). These positions are conserved in sequence among most Pcdh isoforms (Figure S1G) and among classical cadherins (Vester-Christensen et al., 2013), suggesting that these O-glycans play important functional roles. O-mannosylation of cadherins and protocadherins were recently discovered (Vester-Christensen et al., 2013), and it was further shown that O-mannosylation of E-cadherin is essential for preimplantation development of the mouse embryo (Lommel et al., 2013).

The Pcdh structures show local Pcdh-specific embellishments on the EC domain fold. In particular, Pcdh EC1 domains show a number of differences from vertebrate cadherin EC1 domains (Figure S1D), as was previously observed in NMR structures of Pcdhα4 and Pcdhβ14 EC1 domains (Morishita et al., 2006). The A strand is shorter than that of classical cadherins and lacks the conserved Trp-2 residue, which anchors the strand-swap *trans*-binding interface of classical cadherins (Figures S1C and S1D; Posy et al., 2008). The EC1 EF loop region in each of the Pcdh structures contains a disulfide-constrained loop formed by a Pcdh-specific CX_5C motif. The EC2 and EC3 domains of the Pcdh structures are each most similar to either the EC1 or EC2 domain from the atypical cadherin-23 (RMSD 1.5 and 1.2 Å). However, the D and E strands of Pcdh EC2 domains, and the CD loop region of EC3, are significantly longer than found in cadherin-23 or in classical cadherins (Figure S1E). There are also distinctive differences among the structures of the four Pcdh isoforms. The EC1 BC loop helix, C strand, and CD loop regions display distinct conformations in all four structures (Figure S1F). In EC3, the two C-type structures (PcdhαC2 and PcdhγC5) have a longer FG loop than Pcdhβ1 and PcdhγA8, a feature conserved among α and C-type Pcdhs (Figure S1F).

Analysis of the molecular packing of the four Pcdh EC1–EC3 structures revealed different crystallographic contacts for each isoform with no interfaces in common. Interfaces exhibiting typical protein-protein interface attributes were not identified in any of the crystal forms analyzed.

Analytical Ultracentrifugation and Cell Aggregation Assays Define the Multimeric Structure of Pcdhs

We expressed and purified proteins from a C-terminal deletion series comprising EC1–EC6, EC1–EC5, EC1–EC4, and EC1–EC3 and a construct comprising domains EC2–EC6 where EC1 was deleted. Using analytical ultracentrifugation (AUC), we assessed the oligomerization state of each of these ectodomain fragments in solution. With the exception of PcdhγA8, all EC1–EC3 Pcdh isoform fragments behaved as monomers (Table 1). This finding was consistent with our crystal structures in which no apparent binding interfaces were detected. The PcdhγA8 EC1–EC3 fragment formed a disulfide-linked dimer through cysteine 283 in the EC3 domain (Figures S2A and S2B); however, this disulfide bond is likely artifactual since it is not detected in the larger PcdhγA8 isoform fragment (EC1–EC4) (Table 1).

In contrast to monomeric EC1–EC3 fragments, EC1–EC4 or EC1–EC5 Pcdh fragments were observed to self-associate as

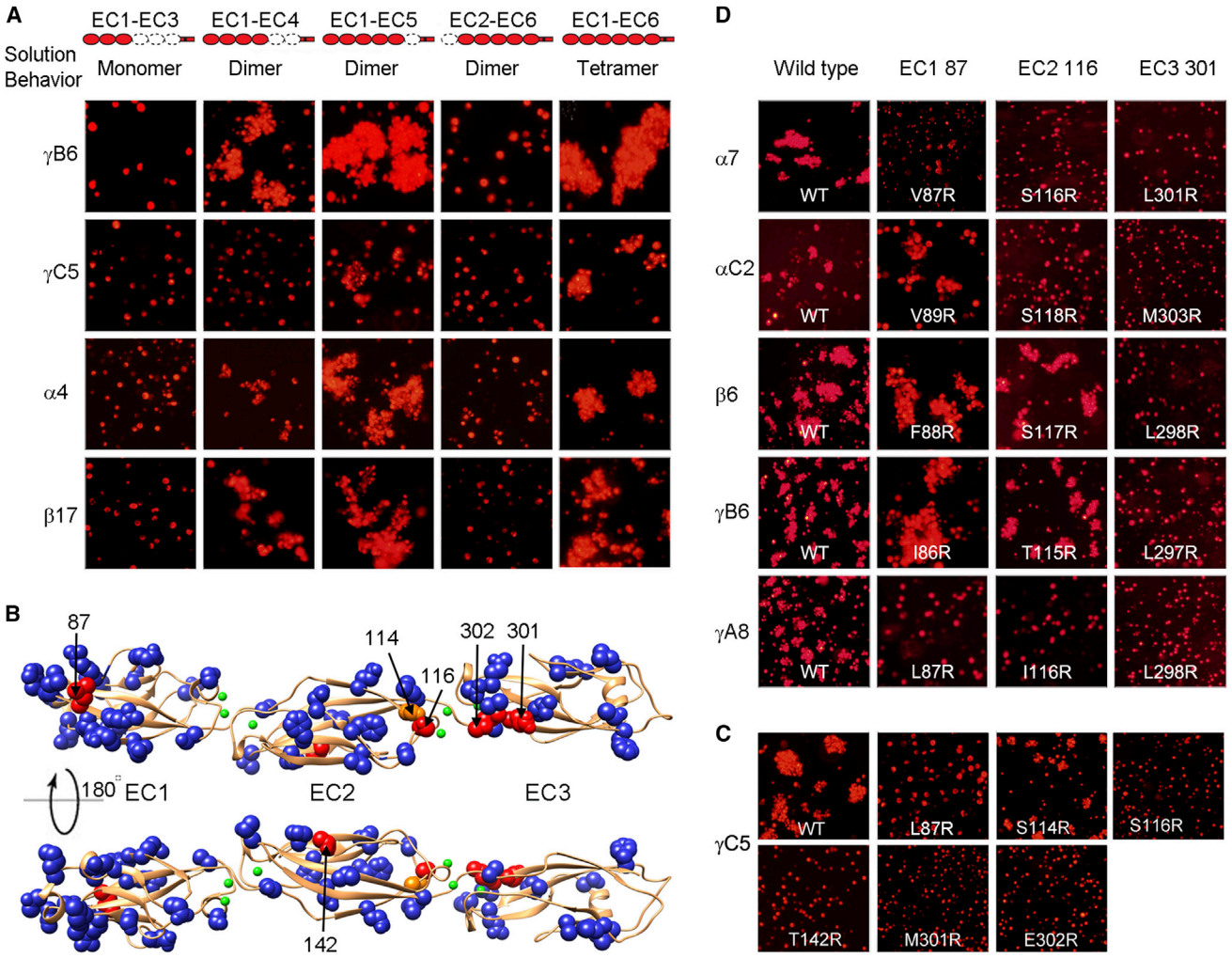


Figure 2. Elements of Pcdh *cis* and *trans* Binding

(A) Correlating multimerization states of truncated Pcdh proteins with their cell-cell recognition properties. Cells transfected with Pcdh deletion series plasmid constructs were tested for aggregation. With the exception of EC2–EC6 Pcdh fragments and PcdhγC5 EC1–EC4, all deletion proteins that formed oligomers in solution also mediated cell aggregation. Full-length Pcdhα4 includes the EC6 domain from PcdhγC3 so it could be delivered to the cell surface.

(B) Probing homophilic interaction interface by arginine-scanning mutagenesis. Residues mutated to arginine are drawn in space filling representation. In blue are mutations that did not disrupt recognition, in orange are mutations that weakened recognition, and in red are mutations that abolished cell-cell recognition. Excluding residue 142, all the effective arginine mutants are located along one side of the molecule.

(C) Cell aggregation experiments showing the mutations in part (B) that weakened or abolished interactions. See also Figure S2C.

(D) In other Pcdh isoforms, residues analogous to the effective PcdhγC5 arginine mutants had similar effects on the cell-cell recognition in the majority of cases.

dimers with dissociation constants (K_D) in the micromolar range (2.9–100 μM) that varied significantly between isoforms (Table 1). The EC1-deleted constructs comprising domains EC2–EC6 also formed homodimers in solution, with K_D values in the low micromolar range (8.9–23 μM). Importantly, AUC measurements for complete ectodomains, including EC1–EC6, could be fit only to a tetramer (dimer-of-dimers) model, indicating a crucial role for the EC6 domain in Pcdh association (Table 1).

We expressed similarly truncated Pcdhs in K562 cells and assessed their ability to mediate cell aggregation. K562 cells provide a robust assay for Pcdh cell-cell recognition, as they do not express endogenous Pcdhs and do not spontaneously aggregate in liquid culture (Reiss et al., 2006; Schreiner and Weiner, 2010; Thu et al., 2014). Cells expressing the EC1–EC3 fragment, which was found to be monomeric in solution, failed to produce cell aggregates (Figure 2A). In contrast, with the exception of PcdhgC5 EC1–EC4, which forms a non-natural disulfide between monomers, cells expressing EC1–EC4, EC1–EC5, or the complete ectodomain (EC1–EC6) showed extensive aggregation for all isoforms tested (Figure 2A). Consistent with previous studies (Schreiner and Weiner, 2010; Thu et al., 2014), cells expressing Pcdh EC2–EC6 fragments, which were shown above to homodimerize in solution, did not aggregate (Figure 2A). Detection of two independent dimers, one of which (generated by EC1–EC4 and EC1–EC5 fragments) correlates with cell-cell aggregation, whereas the other (generated by

EC2–EC6 fragments) does not (Figure 2A), strongly suggests that EC1–EC4 and EC1–EC5 fragments mediate *trans* interactions while the EC2–EC6 fragments mediate *cis* interactions involving the most membrane-proximal domain, EC6 (see also below). The observation that full-length ectodomains form apparent tetramers in AUC strongly suggests that this molecular species corresponds to a dimer-of-dimers formed by these two distinct interfaces, one mediating *cis* and the other *trans* interactions.

Structural Elements of the *trans*-Binding Interface

Arginine-Scanning Mutagenesis

Selected non-basic surface residues of the PcdhγC5 EC1–EC3 domains revealed in the crystal structure were individually mutated to arginine, and the homophilic recognition function of these single-arginine mutant proteins was assessed using the K562 cell aggregation assay. Selected basic surface residues were mutated to glutamic acid. As expected, the majority of single-point mutant proteins exhibited wild-type cell aggregation phenotypes (Figure S2C). In contrast, cells transfected with the arginine point mutant L87R in the EC1 domain, S116R and T142R in the EC2 domain, and M301R and E302R in the EC3 domain of PcdhγC5 showed no detectable aggregation (Figures 2B and 2C). Cells transfected with the EC2 S114R mutation showed diminished homophilic binding (Figure S2C). S114 and S116 are located in the AB loop connecting the A and B β strands in EC2, whereas M301 and E302 are located in the FG loop of EC3. All are located on one side of the molecule and are very close to one another in space, thus defining a potentially continuous homophilic recognition interface with elements distributed over the EC2 and EC3 domains. Notably, L87 in EC1 faces in the same direction although T142 in EC2 does not.

To determine whether this binding region is unique to PcdhγC5, we produced mutants for isoforms from all three Pcdh gene clusters for residues structurally equivalent to PcdhγC5 positions 87, 116, and 301. Mutations equivalent to 301R abolished homophilic recognition for isoforms from all three gene clusters (Pcdhα7, PcdhαC2, Pcdhβ6, PcdhγA8, and PcdhγB6; Figure 2D). Homophilic recognition was abolished for mutations equivalent to 116R for isoforms from the α and γ gene cluster members (Pcdhα7, PcdhαC2, and PcdhγA8), but not for the isoforms we tested from the β and γB clusters (Figure 2D). Finally, mutations equivalent to L87R abolished homophilic recognition for PcdhγA8 and diminished homophilic recognition for Pcdhα7. It is possible that homophilic recognition for the Pcdhβ6 and PcdhγB6 isoforms may not involve residues 87 in EC1 and 116 in EC2, or alternatively, arginine mutants of these residues might not appropriately test their contribution to binding. Below, we show that isoforms from the α and β gene clusters do in fact utilize interface residues in the EC2 AB loop region and others in close structural proximity to EC1 residue 87.

Domain Shuffling to Identify Specificity-Determining Domains

Within each of the mouse gene clusters, there exist pairs of Pcdh isoforms (Pcdhα7 and Pcdhα8; Pcdhβ6 and Pcdhβ8; PcdhγA8 and PcdhγA9) with greater than 80% pairwise sequence identity within their EC1–EC4 domain regions. Despite this high identity, these pairs display strict homophilic specificities (Thu et al., 2014). In order to help identify the binding interface, we produced chimeras in which EC domains were shuffled between the closely related isoforms. These proteins were tagged at the C terminus with either of the fluorescent proteins mCherry or mVenus and tested for binding specificity in the K562 cell assay. We confirmed that all three pairs bind strictly homophilically (Figure 3A, 1–4; Figure 3B,1–4; Figure 3C, 1–4).

The results of cell aggregation experiments using different chimeric constructs are summarized in Figures 3 and S3. These results are presented in such a way that two closely related wild-type "parent" proteins appear at the left of each panel, while each figure indicates whether a particular chimera co-aggregates with one or the other parent protein or prefers to aggregate homophilically. Figure 3D summarizes the data presented in Figures 3A–3C. All chimeric constructs containing EC1–EC3 domains from one isoform and EC4–EC6 domains from another co-aggregated with the wild-type "parent" isoform that contained the same EC1–EC3 domains (Figures 3A–3C, panel 6, and Figures S3B and S3D, panel 13), whereas chimeric constructs with just EC2–EC3 shuffled, preferred to aggregate homophilically (Figures S3A–S3E, panels 11 and 12).

Despite the fact that shuffling EC1–EC3 is sufficient to swap specificity in close pairs, our AUC and cell aggregation assay results (Table 1 and Figure 2A) indicate that all four N-terminal domains (EC1–EC4) are required for *trans* homophilic recognition. We therefore generated a chimera of PcdhγA8 in which domains EC2–EC4 were replaced with the corresponding domains of the closely related PcdhγA9 isoform, while domains EC5–EC6 were replaced with the EC5–EC6 domains of the distant PcdhγB6 isoform, which would not be expected to interact in *trans* with PcdhγA8 or PcdhγA9. Cells expressing this chimera adhere to cells expressing PcdhγA9 indicating, consistent with AUC data, that the EC4 domain plays a role in determining homophilic binding specificity (Figure 3C, panel 8). This conclusion is also supported by cell aggregation studies using chimeras where EC1 is derived from one parent and EC2–EC6 from another. In all cases, these chimeras co-aggregate with the parent containing the same EC2–EC6 domains (Figure S3A, S3C, and S3E, panel 1; Figures S3B and S3D panel 2). Since domains EC5 and EC6 are not required for *trans* binding, these results also implicate EC2–EC4 as sufficient to determine homophilic specificity.

The experiments reported in Figure S3 help define the minimal number of domains within the EC1–EC4 region that determine the binding properties of a chimera. The presence of a single domain is never enough to mediate co-aggregation with a parent isoform containing this domain (Figure S3A, S3C, and S3E, panels 2, 4, and 6; Figures S3B and S3D, panels 1, 3, and 5), but in some cases, a mismatched single domain is capable of disrupting binding to the parent isoforms (Figure S3C, panel 5; Figure S3D, panel 6; Figure S3E, panel 3). In a few cases, the presence of just two domains in common is sufficient to mediate co-aggregation with a parent even if the other four domains are different. This can be seen in a chimera containing EC1 and EC3 from γA9 and EC2 and EC4–EC6 from γA8, which co-aggregates with wild-type γA9 (Figure S3C, panel 10), and a chimera

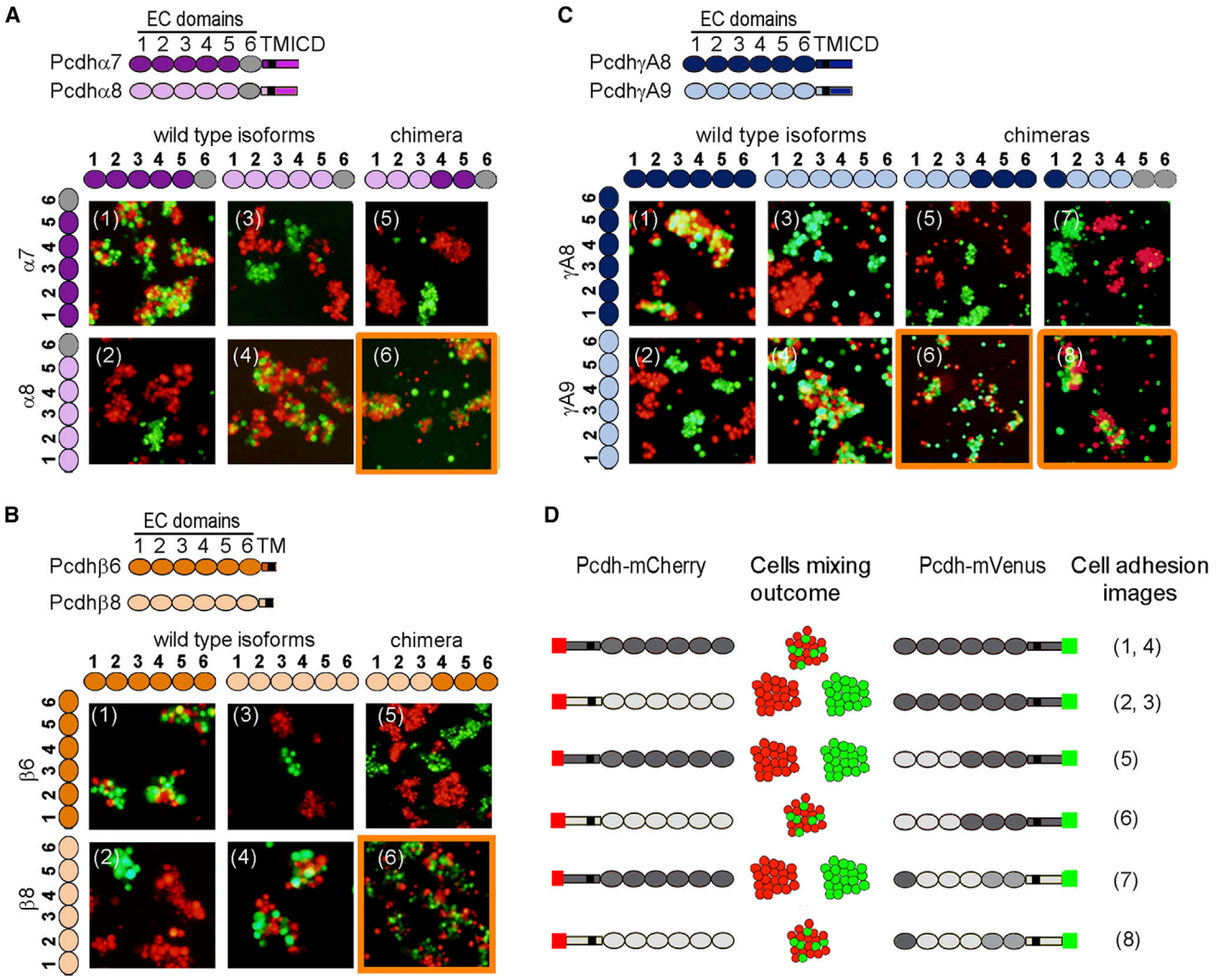


Figure 3. Pcdh *trans* Binding Depends on the Four N-Terminal Domains EC1–EC4

(A–C) Domain-shuffled chimeras of closely related isoforms and their wild-type counterparts were assayed for binding specificity. Swapped specificity was noted for chimeras in which either the EC1–EC3 or EC2–EC4 domains were replaced with the corresponding domains of closely related isoforms. See also Figure S3.

(D) Schematic representation of the domain-shuffled isoforms and their observed binding specificities to their wild-type isoform counterparts.

containing EC1 and EC2 from β8 and EC3–EC6 from β6, which co-aggregates with wild-type β8 (Figure S3E, panel 8). Overall, these results are consistent with all four N-terminal domains, EC1–EC4, contributing to *trans* binding with the relative contributions of each domain to specificity varying from one isoform to another.

Rational Design of Point Mutations to Identify Specificity-Determining Residues

Sequence alignment of specificity-determining EC3 domains shows that Pcdhα7 and Pcdhα8 differ in five amino acids, whereas PcdhγA8 and PcdhγA9 differ in eight (Figure 4A). Notably, in both cases, three of these residues are located in the same structural element: the FG loop (Figures 4A, 5A, and 5C). In the case of PcdhγA8 and PcdhγA9 the three variable FG loop residues are highly conserved within their respective orthologs (Figure 4B). Together, these data strongly suggest that these three EC3 domain FG loop residues act as specificity determinants for α and γ Pcdh isoforms.

To test this hypothesis experimentally, we swapped the three residues (Figure 5) between the EC3 domains of closely related isoforms and tested their binding specificities with their "parent" native isoforms. We produced chimeras with the three FG-loop residues of one isoform replaced with the corresponding residues of its close-pair isoform. These three-residue-swapped mutants were tested, along with their native "parents," in the K562 cell aggregation assay. Cells expressing an isoform in which the three FG-loop residues were replaced with those from the close-pair isoform intermixed with cells expressing the wild-type isoform with residues identical to those at the shuffled positions (Figures 5A and 5C). In contrast, these cells segregated from cells expressing the wild-type isoform from which the EC3 domain originated (Figure S4). We conclude that the three

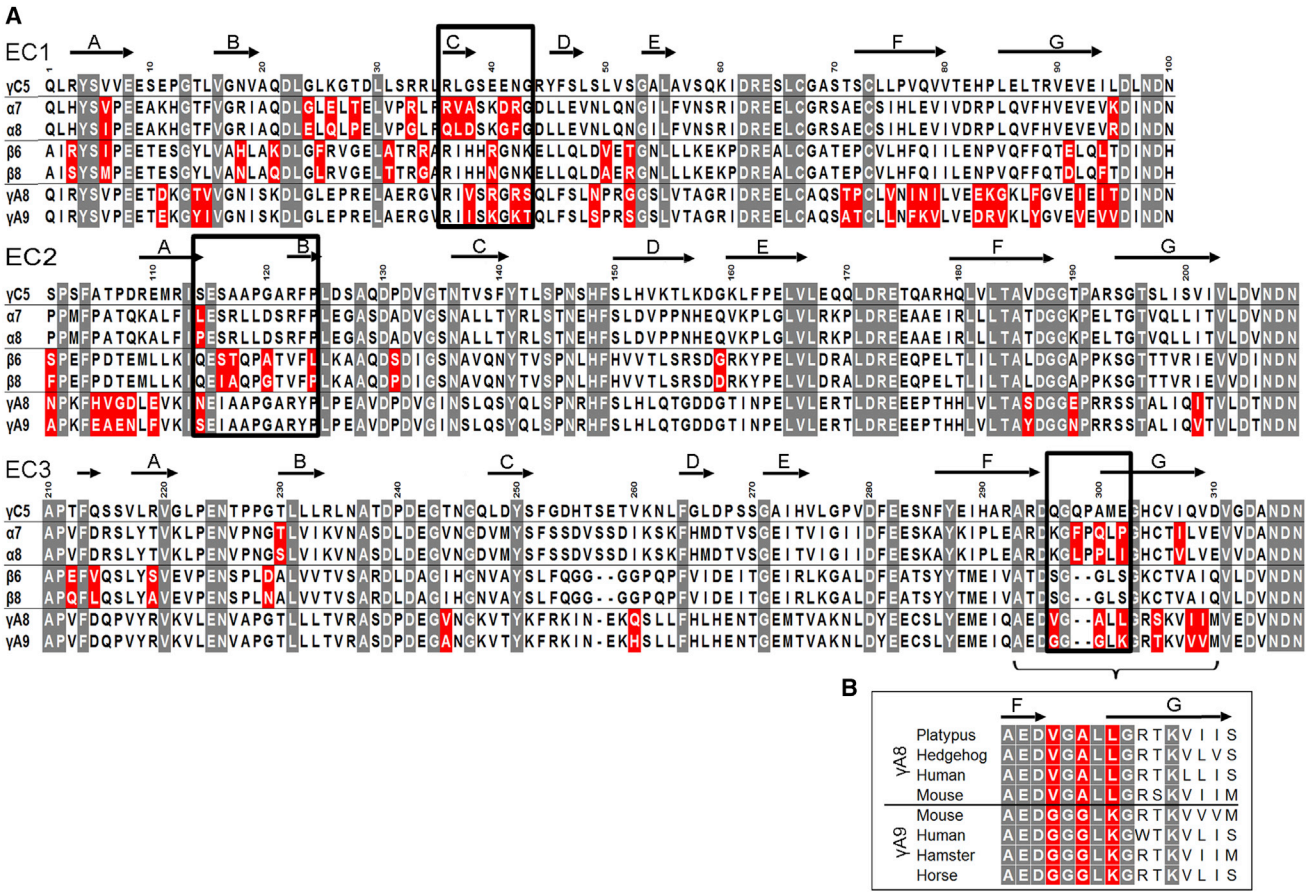


Figure 4. Candidate Specificity Determining Residues

(A) Multiple sequence alignment of the three closely related Pcdh isoform pairs, along with PcdhγC5. Highlighted in gray are positions conserved in all Pcdh sequences. Sequence positions that differ between the closely related isoforms are shown in red; a subset of these residues determines binding specificity. Residues swapped between isoforms and assayed for binding properties are boxed. Secondary structure from PcdhγC5 is shown at the top of the alignment. (B) Multiple sequence alignment of the FG-loop region for PcdhγA8 and PcdhγA9 orthologs. Three of the residues that differ between mouse PcdhγA8 and PcdhγA9 are highly conserved in orthologs (highlighted in red), suggesting their functional importance.

variable residues of the EC3 FG loop are specificity determining in the closely related α and γ isoforms.

A similar analysis was carried out for EC1 and EC2 domains with comparable results. As with the EC3 domains, we analyzed close isoform pairs (Figure 4A) and identified candidate specificity-determining residues located on the EC1 C strand and EC2 AB region (Figure 5). We validated these assignments by showing that shuffling residues between EC2 domain AB regions resulted in swapped specificities for close-pair isoforms from all three *Pcdh* gene clusters (Figures 5 and S4). Shuffling residues between EC1 domain C strand regions was sufficient to swap EC1 specificities from Pcdhβ6 to that of Pcdhβ8 or from Pcdhα7 to Pcdhα8. The contribution of this region in the Pcdhγ pair could not be determined because shuffling of residues in this region resulted in a protein that could not mediate cell aggregation (Figure S4D). We note that swapping EC1 specificities from Pcdhβ6 to Pcdhβ8 or EC2 specificities from Pcdhα7 to Pcdhα8 or from PcdhγA9 to PcdhγA8 required the alteration of only a single residue (residue R41N, L114P, and S114N for β, α, and γ respectively; Figure 5).

Rational and Random Mutagenesis Identify the Same Functional Binding Surfaces

Figures 2 and 5 list specificity-determining residues identified from arginine scanning and bioinformatics-based mutagenesis. The finding that two different approaches implicate the same structural regions in Pcdh homophilic binding and that these regions are in common for isoforms from different *Pcdh* gene clusters indicates that these regions—the EC1 C and G strands, the EC2 AB loop, and EC3 FG loop (Figure 5D)—are likely to contribute to determining the binding specificities for other Pcdh isoforms as well. As shown above, EC4 contributes to the *trans* binding specificity in a similar way to that of EC1. However, we focused on the EC1–EC3 domains because this is the region for which we have atomic-level structures.

AUC Experiments on Mutant Proteins Confirm that Pcdh trans Interactions Occur via EC1–EC4 Domains, whereas cis Interactions Occur via the EC6 Domain

We have provided evidence from both AUC and cell aggregation assays that the EC1–EC4 domains mediate Pcdh *trans* interactions, whereas the EC6 domain mediates an independent Pcdh

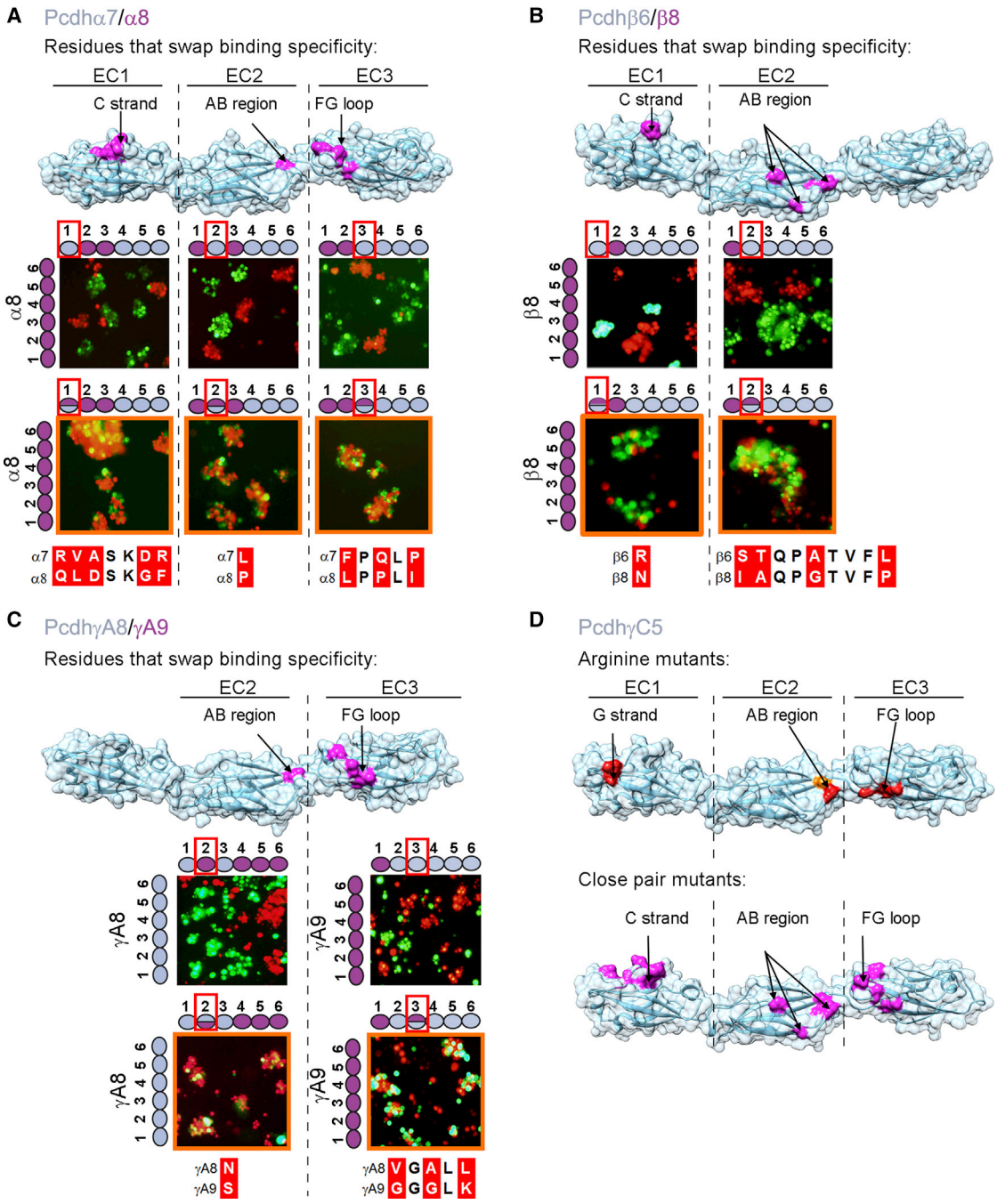


Figure 5. Structural Elements of the Canonical Pcdh *trans* Binding Interface

(A–C) Assessing specificity-determining residues. Binding properties of wild-type isoforms (left side of each panel) or constructs with shuffled residues (top of each panel) were tested separately for each EC domain. Cases in which shuffled residues swapped specificities are indicated by an orange outline. Residues shuffled between closely related isoforms are shown in magenta on surface representations of the Pcdhα7, Pcdhβ6, and PcdhγA8 structures. Sequence alignments of shuffled regions are shown. See also Figure S4.

(D) Correspondence between *trans* interface residues identified by arginine scanning and close-isoform pair analysis. Single arginine mutant residues that abolish or diminish homophilic binding, highlighted in red and orange respectively, are found in the same structural regions as the shuffled residues (see also Figure 2). Residues that swap binding specificity between closely related isoforms are shown in magenta on surface representations of the Pcdh-γC5 crystal structure.

cis interaction. To provide further evidence for these findings, we expressed and purified various domain-truncated constructs of PcdhγA8-I116R, PcdhγC5-S116R, and PcdhαC2-S118R. Since an arginine at these positions ablates *trans* binding in cell aggregation assays, these mutant constructs should only affect the Pcdh *trans*-association but not the *cis*-association in AUC

experiments. As expected, the EC1–EC4 fragment of I116R PcdhγA8 behaved differently from its wild-type counterpart and was monomeric in solution (Table 1). In contrast, we found that, similar to its wild-type counterpart, the EC2–EC6 fragment PcdhγC5-S116R behaved as a dimer with K_D similar to wild-type EC2–EC6. This observation suggests that the EC2–EC6 protein dimerizes in *cis* through a region that is not involved in the *trans* interface (Table 1). Finally, the complete ectodomain of PcdhαC2 containing an S118R mutation displayed tetramerization affinity, which was an order of magnitude lower than that of the wild-type protein. Similarly, the S116R mutant of PcdhγC5 EC1–EC6 did not form tetramers (as does its wild-type counterpart) but rather, similar to the EC2–EC6 fragment, self-associates as a dimer. Since *trans* binding has been ablated by this mutation, the observed dimer must correspond to association in *cis* (Table 1).

The *trans* Homophilic Interface Is Formed via Head-to-Tail Interactions of EC1–EC4 Domains

Computational Docking Yields Antiparallel Orientations

We carried out modeling studies in an effort to elucidate the dimerization mode of Pcdhs. We limited our modeling to EC1–EC3, for which we have determined crystal structures and have identified specificity-determining residues. We used the M-zdock program (Pierce et al., 2005) to produce symmetric homodimeric models for the EC1–EC3 domain regions of PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5. We generated thousands of models for each crystal structure and used the experimentally identified specificity determinant residues to filter the docked models; requiring models to include these residues at the binding interface. A second constraint required docking models to have a buried surface area at the binding interface of more than 1,200 Å^2 (600 Å^2 per protomer). Applying these two conditions reduces the number of docked models from thousands to 149: 23, 40, 40, and 46 for PcdhγA8, Pcdhβ1, PcdhαC2, and PcdhγC5, respectively. We then structurally clustered the filtered docked homodimers with the expectation that there would be more docked structures near the native conformation.

Notably, the majority of the filtered docked homodimeric Pcdhs (62.5%) adopted a head-to-tail orientation of the two molecules in which the EC2 domain of one molecule interacts with the EC3 domain of its partner (Figures 6A and S5A, i and ii). Furthermore, most structures with this binding mode place the EC1 domain of one molecule adjacent to the expected position of the EC4 domain of its partner (Figure 6A). Only three of the docked and filtered complexes had a head-to-head orientation (two for PcdhγC5 and one for PcdhαC2; Figure S5A, iii), whereas filtered solutions for Pcdhβ1 and PcdhγA8 resulted solely in solutions with a head-to-tail orientation. We note that it is the application of the two constraints, one of which was experimentally derived, that results in this distribution of binding modes.

Experimental Validation of a Head-to-Tail Orientation

The computational evidence for a head-to-tail dimer, taken together with our identification of EC1–EC4 as the specificity-determining region, suggests that EC1 interacts with EC4 and EC2 interacts with EC3. In order to validate this model, we carried out cell aggregation assays on chimeras of the γA8 and γA9 Pcdh isoforms, which were designed to determine which domains physically interact. As shown in the schematic, diagrams in Figure 6B (panels 1–3), head-to-tail binding would result in a dimer where all EC2/EC3 and EC1/EC4 interactions involve domains from the same wild-type protein. In all three cases, the chimeras form mixed aggregates, thus providing strong evidence for our proposed model of the Pcdh-Pcdh interface. Note that, if the monomers bound in a head-to-head orientation, some interacting domains would be derived from different wild-type proteins so that mixed aggregates would not be expected to form.

Figure 6B (panels 4 and 5) provides direct evidence that EC1 interacts with EC4 and EC2 interacts with EC3. Comparing panel 4 to panel 1, the only difference between the two is that there is a mismatch between EC4 and EC1 in panel 4. The two cell populations in panel 4 form separate aggregates, indicating that this single mismatch is sufficient to ablate *trans* dimerization. An identical conclusion regarding EC2 and EC3 is reached by comparison of panel 5 to panel 2. Here again, a single-domain mismatch inhibits co-aggregation even though the remaining three domains are correctly matched.

To further validate the model of head-to-tail binding, we carried out mutagenesis experiments on specificity-determining regions. Since, as shown above, for the α and γ close pairs the EC2 AB loop and the EC3 FG loop determine specificities, we reasoned that the specificity-determining residues in the EC2 AB loop might interact with corresponding residues in the EC3 FG loop. Notably, the largest cluster of structurally similar docked and filtered complexes is the only cluster that positions the EC2 AB loop near the EC3 FG loop and projected to position the EC1 near EC4 (Figures 6A and S5A). To test this model (Figure 6A), we relied on two observations (1) that arginine mutations of residue 301 in the EC3 FG loop region and residue 116 in the EC2 AB loop region (PcdhγC5 numbering) abrogate recognition in isoforms from different gene clusters (Figures 2B–2D) and (2) that docked models position residue 301 and residue 116 at close distance (less than 6Å, Figure 6A). Hypothesizing that residues 116 and 301 are near each other in the recognition complex, we attempted to rescue single-arginine mutants at residue 303 of PcdhαC2 or 298 of PcdhγA8 and Pcdhβ6 (analogous to PcdhγC5 301) by producing an aspartic acid mutation of PcdhαC2 residue 118, of PcdhγA8 residue 116 or of Pcdhβ6 residue 117 (analogous to PcdhγC5 116). The designed double mutants could, in principle, form a salt bridge at the interface and thus might rescue recognition.

For all three isoforms (PcdhαC2, Pcdhβ6, and PcdhγA8), cells expressing the double arginine/aspartic-acid mutants tested positive for cell aggregation (Figure 6C), indicating that these two mutated residues (116 and 301), located respectively on domains EC2 and EC3, are in close proximity at the homophilic binding interface. This observation provides strong support for a head-to-tail binding mode where EC2 interacts with EC3 and where EC1 interacts with EC4. Moreover, since PcdhαC2, Pcdhβ6, and PcdhγA8 are not closely related, it is likely that the modeled interface represents the recognition interface for other Pcdhs as well.

DISCUSSION

Counterintuitively, the phenomenon of neuronal self-avoidance is initiated by *trans* homophilic adhesive binding between Pcdhs.

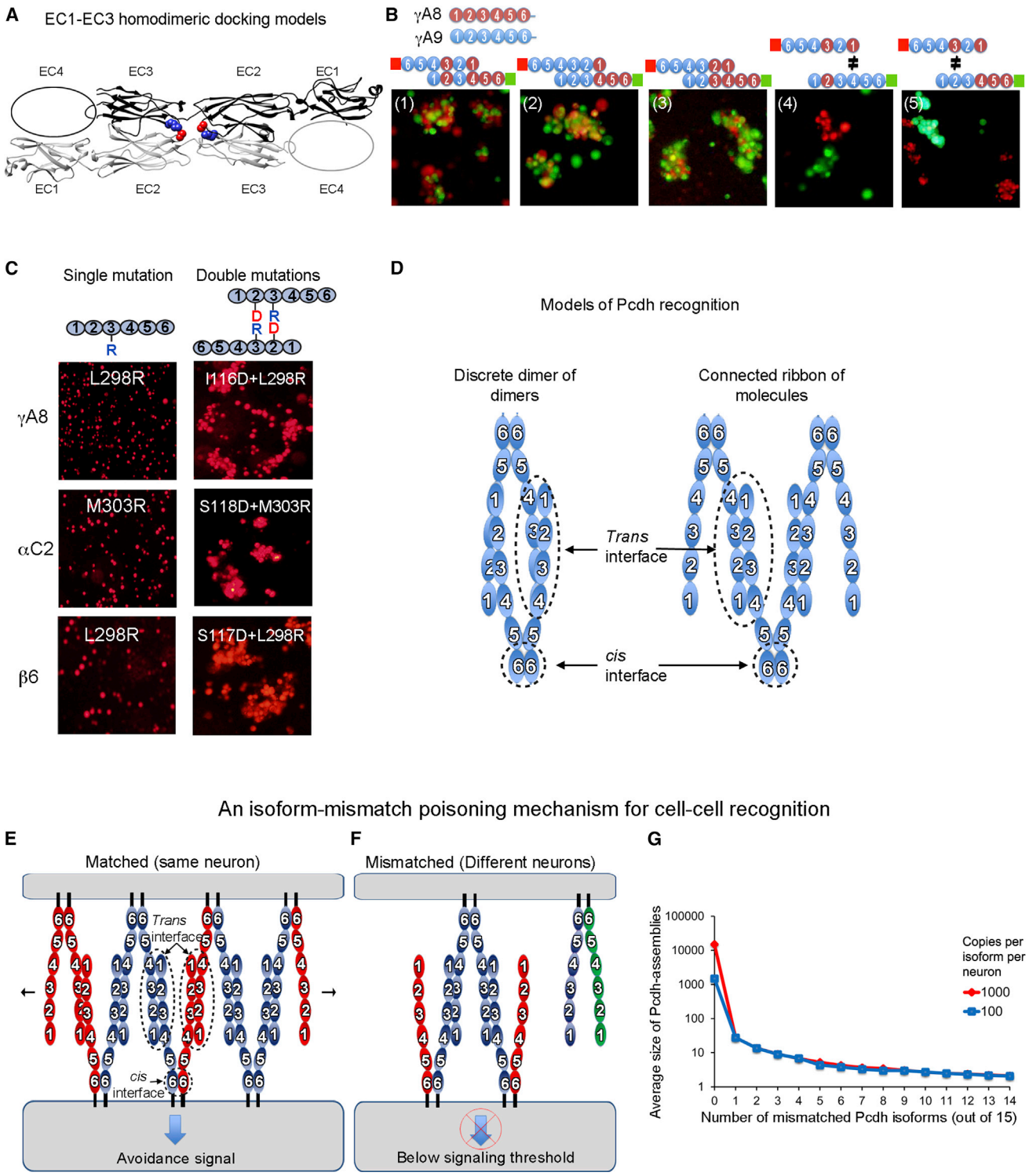


Figure 6. Molecular Logic of Pcdh-Mediated Cell-Cell Recognition

(A) Shown in ribbon representation is the only orientation observed for docking of the four EC1–EC3 domains structures, which position the EC2 AB loop in close proximity to the EC3 FG loop. EC2 AB loop residue 116 and FG loop residue 301 are drawn as space filling and colored red and blue, respectively. The vast majority of the docked complexes were observed to interact in this mode. See also Figure S5A.

(B) Cell aggregation assays on chimeric proteins that show EC1 interacts with EC4 and EC2 interacts with EC3. Schematic representation of the head-to-tail interaction between the domain-shuffled chimeras is shown above each panel. Mixed aggregates were formed where all interactions involve "matching" domains (panels 1–3). Separate aggregates were formed when there is a mismatch between EC1/EC4 (panel 4) or between EC2/EC3 (panel 5).

(legend continued on next page)

Presumably, repulsion is a consequence of the activation of downstream signals via the ICD, which is known to interact with signaling adaptors and kinases (Han et al., 2010; Schalm et al., 2010). This mechanism requires that different neurons express a sufficiently distinct set of Pcdh isoforms so that inappropriate ''self''-recognition, and subsequent repulsion, will not occur. In the case of invertebrates, this is accomplished through the stochastic expression of about 10–50 different alternatively spliced Dscam isoforms in each cell (Hattori et al., 2008; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010). With thousands of stochastically generated distinct Dscam isoforms, the probability that two different neurons express the same set of isoforms is extremely low (Miura et al., 2013). Considering the much smaller number of distinct Pcdh isoforms in vertebrates, isoform diversity alone cannot account for ''non-self-discrimination.''

As mentioned above, we have shown previously that an interference phenomenon plays a crucial role in Pcdh-based non-self-discrimination (Thu et al., 2014). In this paper, we present evidence from several independent sources of data that suggest that Pcdh cell-cell recognition is mediated by a mechanism that couples *cis* and *trans* interactions. Specifically, we propose that Pcdh isoforms form promiscuous EC6 dependent *cis*-dimers at the cell surface that associate specifically in *trans* via a stereotyped interface with elements in domains EC1–EC4. Below, we summarize our findings and discuss their implications for the molecular mechanisms by which clustered Pcdhs mediate neuronal self-recognition and non-self-discrimination.

Pcdh Homophilic Specificity Is Determined by a Head-to-Tail *trans* Recognition Interface

We found that Pcdh EC1–EC3 fragments do not associate in solution, nor do they mediate homophilic cell-cell recognition in cell aggregation assays. Rather, we showed both in AUC measurements and cell assays that stable *trans* dimerization requires all four of the N-terminal EC1–EC4 domains. Site-directed arginine scanning mutagenesis and rational mutagenesis based on analysis of sequence alignments allowed us to identify key structural elements in a *trans* interface that mediate cell-cell recognition between Pcdhs.

The identification of interfacial regions in EC2 and EC3 through computational modeling and mutagenesis experiments provided strong constraints that made it possible to demonstrate that Pcdh *trans* dimers adopt a head-to-tail orientation where EC2 interacts with EC3. This remarkable anti-parallel *trans*-interaction is in contrast to the parallel *trans* dimerization of classical cadherins. However, for classical cadherins, the parallel binding mode is made possible by a significant intramolecular bend whereby the five EC domains form a highly curved structure so that interacting membrane-distal EC1 domains from apposed cells are parallel to one another. In contrast, since the EC1–EC3 domains in Pcdhs are straight rather than curved, binding in parallel would require a sharp bend between the three N-terminal and three C-terminal domains. Such a bend has been observed only in cadherins lacking inter-domain calcium binding sites (e.g., DN cadherin [Jin et al., 2012]), and the presence of complete calcium binding sites between all domains renders such significant bending highly unlikely in the case of Pcdhs.

Figure 6A shows the structure of an EC1–EC3 *trans* dimer obtained from our docking studies that satisfies all the constraints established by mutagenesis. The EC4 domain is represented as an ellipse in the diagram since its structure has not yet been determined. In addition to satisfying all the mutagenesis data used as constraints in the docking studies, independent evidence supporting the model includes (1) the set of five cell aggregation studies on γA8 and γA9 chimeras (Figure 6B) that show that EC1 interacts with EC4 and EC2 interacts with EC3 and (2) the rescue experiments shown in Figure 6C that reveal that residue 116 in EC2 is in close proximity to residue 301 in EC3, as predicted by the head-to-tail model (Figure 6A).

The head-to-tail model shown in the figure provides a clear explanation of the binding affinity and cell aggregation data. In the model, the free energy of binding is distributed over all four domain-domain interfaces, and all must be present to generate sufficient affinity to produce a stable homodimer. This is evident from the observations that three domain constructs do not dimerize and that interfacial mutations in only a single domain are sufficient to ablate binding. All EC1–EC3 ectodomain fragments studied here were monomeric, and none revealed a likely *trans* interaction. With a head-to-tail orientation, deletion of only one domain in EC1–EC4 effectively removes half the interface, providing a likely explanation for the absence of native dimer interactions.

We note that the structural model itself is unlikely to be accurate in detail and will certainly be superseded once X-ray structures of all four interacting domains are available. The major significance of the model is the demonstration that Pcdhs dimerize in *trans* in a head-to-tail orientation with an extended interface formed from four inter-domain interfaces (two EC2/EC3 and two EC1/EC4).

(C) The EC2 domain AB region recognizes the EC3 domain FG loop. Cells expressing isoforms with single arginine mutants in the EC3 FG loop region or with double mutations (aspartate at the AB region and arginine at the FG loop) were assayed for aggregation. The double mutation rescued the non-adhesive phenotype, supporting the head-to-tail binding orientation shown in part (A).

(D) Two possible models of Pcdh interaction. A discrete tetramer composed of a dimer of dimers is observed in analytical ultracentrifugation, but we suggest that a connected ribbon of molecules can form between cells via the *trans* and *cis* interactions.

(E and F) A model for Pcdh-mediated cell-cell recognition based on formation of a superstructure defined by promiscuous *cis* and specific *trans* interactions. Growth of the chain of molecules requires matching of all isoforms; a single mismatch can terminate chain extension. Dendrites of the same neuron will have the same isoform repertoire, whereas dendrites of different neurons will differ. In this model, repulsion signaling is triggered, or achieves a sufficient level for response, only through the formation of an extended chain of Pcdhs.

(G) For the case of 15 distinct Pcdh isoforms expressed per cell, Monte-Carlo simulations were used to estimate the average size of one-dimensional Pcdh assemblies between contacting cells. The average number of *cis* dimers that comprise such assemblies is shown on a logarithmic scale as a function of the number of mismatched isoforms. Two cases are shown—one for 15,000 total Pcdh monomers (1,000 per isoform, red), and one for 1,500 total copies (100 per isoform). The model assumes that each cell contains a stable set of *cis* dimers formed from the random association of monomers present in each cell. See also Figure S5B.

We note that the molecular dimerization logic of Pcdhs, where different domains recognize one another through EC1/EC4 and EC2/EC3 *trans* interactions, is fundamentally different from that of Dscam1, where the dimerization interface is formed from three separate self-self-interactions, Ig2/Ig2, Ig3/Ig3, and Ig7/Ig7.

Pcdhs Form *cis* Dimers Mediated by EC6

We previously provided evidence for promiscuous Pcdh EC6/EC6 *cis* interactions. Specifically, any single carrier isoform (β, γ, or C-type) can mediate cell-surface delivery of α isoforms, which are otherwise confined within the cell, through interactions involving the EC6 domain (Thu et al., 2014). In addition, the pairwise sequence identity between EC6 domains for all isoforms of Pcdhβ or Pcdhγ clusters averages over 90% (Thu et al., 2014), which is consistent with the idea of promiscuous interactions.

We show above that the EC6 domain mediates Pcdh *cis* dimerization even in the absence of *trans* interactions. Moreover, as shown in Table 1, the affinity of this interaction is comparable or even stronger than the *trans* interaction involving EC1–EC4. In general, *cis* interactions in the two-dimensional environment of the plasma membrane would be significantly enhanced, and the effect is strongest for membrane proximal domains, as there would be little entropy loss due to inter-domain flexibility upon binding (Wu et al., 2011, 2013). Indeed, even at low surface densities, molecules with substantial solution (3D) K_Ds, such as that of Pcdhs, will likely form dimers on cell surfaces. The promiscuity of the EC6 carrier function suggests that these dimers can form between essentially any two Pcdh isoforms, which in turn suggests that Pcdhs on cell surfaces exist as *cis* dimers formed by pairs of different isoforms from all three subfamilies as well the C-type isoforms.

Assembly Termination by Mismatched Isoforms Distinguishes Self from Non-self

We have shown above that full-length Pcdh ectodomains in solution form tetramers (a *cis*/*trans* dimer of dimers) mediated by head-to-tail *trans* interactions involving EC1–EC4 and a *cis* interaction involving EC6. A schematic of this molecular arrangement is shown in the left panel of Figure 6D. If Pcdhs on cell surfaces interacted in this manner, cellular recognition would be based on dimeric recognition units. However, as we have discussed in a previous study, dimeric recognition units are unlikely to provide sufficient diversity for neuronal non-self-discrimination, and indeed all models based on multimeric recognition units encounter difficulties in accounting for both self-recognition and non-self-discrimination (Thu et al., 2014). For this reason, we previously proposed an alternative recognition mechanism based on "junction-like" molecular assemblies at least partially reminiscent of those formed by classical cadherins.

As discussed above, each Pcdh molecule forms strong independent *trans* and *cis* interactions. This is in contrast to classical cadherins in which each molecule forms relatively strong *trans* interactions and two weak asymmetrical *cis* interactions that become stronger on cell surfaces only once the *trans* interactions have been formed (Wu et al., 2011). In the case of classical cadherins, the combination of *cis* and *trans* interactions generates a two-dimensional lattice that corresponds to the extracellular structure of adherens junctions (Harrison et al., 2011). In contrast, the interactions defined here for Pcdhs suggest the formation of a one-dimensional zipper-like structure involving symmetrical *cis* and *trans* interactions. This structure is depicted in the right panel of Figure 6D, which shows how each bivalent Pcdh *cis* dimer could recognize two other dimers via independent *trans* interactions so as to form a connected ribbon of molecules that emanate from two apposed cell surfaces. We note that still-undiscovered extracellular, trans-membrane, or cytoplasmic interactions may ultimately reveal a more complex network of interactions than the one depicted in the figure. For example, the receptor tyrosine kinase Ret has been shown to associate with, and directly or indirectly phosphorylate, Pcdhα and γ tyrosine residues in their ICDs (Schalm et al., 2010). In any case, the existence of even a one-dimensional network would provide a mechanism for interference that does not encounter the problems based on models of isolated multimeric recognition units.

Figure 6E illustrates that cells with the same isoform composition would be able to form a large assembly upon contact. In contrast, cells with different isoform compositions would incorporate mismatches, preventing further growth of the lattice (Figure 6F). If downstream signaling leading to neurite repulsion depends on the size of the assembly, which in turn depends on isoform composition, the model offers a natural mechanism for Pcdh interference. Indeed, there is a striking dependency of the size of Pcdh assemblies on the number of mismatched Pcdh isoforms. Figure 6G plots the average size of such linear assemblies as a function of the number of mismatched isoforms between two contacting neurons. Assembly size is obtained from Monte-Carlo calculations based on a model that assumes that each cell contains a stable set of *cis* dimers formed from the random association of monomers present in each cell. When all isoforms are identical, assembly size is limited solely by the number of copies of each isoform. Remarkably, the presence of even a single mismatched isoform is sufficient to reduce the average size of an assembly by at least two orders of magnitude. The results presented in Figure 6G thus suggest that a mechanism based on mismatched-isoform chain termination of a linear Pcdh-assembly could provide a binary definition of self and non-self.

While we recognize that this isoform mismatch chain-termination model is speculative, it is consistent with the presence of strong independent *cis* and *trans* interactions. Such signaling systems have been observed previously, including the one-dimensional network of CTLA-4/B7 immune receptors (Schwartz et al., 2001), where signaling has also been proposed to be based on large cell-surface assemblies. Most importantly, the model provides a mechanism whereby 58 Pcdhs can generate the high level of diversity sufficient to allow for neuronal self-avoidance without encountering the problems for self-recognition, which is implicit in previous models that depend on discrete combinatorial multimeric recognition units.

EXPERIMENTAL PROCEDURES

Protein Production and Crystallography

Proteins for crystallization or biophysical analysis were expressed in suspension-adapted HEK293 Freestyle cells (Invitrogen) and purified by nickel affinity and size exclusion chromatography. Pcdh crystals were grown by vapor diffusion in 1–2 μl hanging drops, except the Pcdhβ1 EC1–EC3 crystals, which were grown in 0.2 μl sitting drops. The PcdhγC5 EC1–EC3 $P4_32_12$ crystal

structure was solved using the MIRAS technique, while all the other Pcdh crystal structures were solved by molecular replacement. See the Supplemental Experimental Procedures for details.

Cell Aggregation Assays

Pcdh expression constructs were transfected into K562 cells by electroporation. The transfected cells were grown in culture for 24 hr. Cells were then allowed to aggregate for 1 to 3 hr on a rocker inside an incubator at 37°C. The cells were then fixed in 4% PFA for 10 min, washed in PBS, and cleared with 50% glycerol for imaging. See the Supplemental Experimental Procedures for details.

Sedimentation Equilibrium Analytical Ultracentrifugation

Proteins were diluted to an absorbance at 10 mm path length and 280 nm of 0.65, 0.43, and 0.23 absorbance units. All samples were run at four speeds: 11,000, 14,000, 17,000, and 20,000 rpm (all EC1–EC3 constructs) or 9,000, 11,000, 13,000 and 15,000 rpm (all EC1–EC4, EC1–EC5, and EC1–EC6 constructs), respectively. Measurements were carried out at 25°C and detection was by UV at 280 nm.

Monte-Carlo Simulations

A stochastic algorithm was used to estimate the average size of Pcdh assemblies (number of linked *cis* dimers) formed between a pair of neurons each expressing 15 distinct isoforms with 0–15 common isoforms. It was assumed that a neuron expresses an equal number of copies of each of the 15 Pcdh isoforms, with either 1,000 or 100 copies per isoform (i.e., 15,000 or 1,500 total Pcdh monomers respectively). 10^6 simulations were performed, and in each simulation, stable *cis* dimers were randomly and independently generated for the contacting neurons. Note that the distribution of *cis* dimers on both neurons will not in general be identical even for neurons with an identical set of monomers. A linear network was initiated by randomly choosing a dimer on one of the cells. In the next step, a *cis* dimer is chosen on the second cell where one of its monomer constituents matches one of the monomers in the dimer chosen on the first cell. This matching process is then repeated with the search for matching dimers alternating between the contacting neurons moving from one cell to the other as the chain extends in two directions. This extension process was repeated until there remained no matching dimers either due to a mismatch or to a depletion of dimers.

ACCESSION NUMBERS

The coordinates and structure factors for the reported crystal structures are deposited in the Protein Data Bank under accession codes PDB: 4ZPO, 4ZPQ, 4ZPP, 4ZPN, 4ZPM, 4ZPL, and 4ZPS.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.026.

AUTHOR CONTRIBUTIONS

R.R., C.A.T., K.M.G., T.M., L.S., and B.H. designed research, analyzed data, and assembled and wrote the paper. R.R. carried out the computational analysis. C.A.T. carried out cell aggregation assays and analysis. F.B., S.M., H.N.W., and K.M.G. prepared and crystallized all proteins. H.N.W. and K.M.G. determined the crystal structures. C.A.T., S.M., H.N.W., and M.C. prepared Pcdh mutants. G.A. performed and analyzed the AUC experiments. A.H. and H.C. performed the glycosylation analysis.

ACKNOWLEDGMENTS

We thank Dr. David Hirsh for valuable comments on the manuscript. We thank Igor Kourinov, Surajit Banerjee, and Narayanasami Sukumar at Advanced Photon Source (APS) for support with synchrotron data collection. This work was supported by the National Science Foundation grant to B.H. (MCB-1412472), National Institutes of Health grant to L.S. (R01GM062270), joint NIH grant to T.M. and L.S. (R01GM107571), NIH Training Programs to H.N.W. (T32GM008281), and Danish National Research Foundation (DNRF107) to A.H. and H.C.

Received: May 14, 2015
Revised: July 17, 2015
Accepted: August 27, 2015
Published: October 15, 2015

REFERENCES

Boggon, T.J., Murray, J., Chappuis-Flament, S., Wong, E., Gumbiner, B.M., and Shapiro, L. (2002). C-cadherin ectodomain structure and implications for cell adhesion mechanisms. Science *296*, 1308–1313.

Chen, W.V., and Maniatis, T. (2013). Clustered protocadherins. Development *140*, 3297–3302.

Chen, W.V., Alvarez, F.J., Lefebvre, J.L., Friedman, B., Nwakeze, C., Geiman, E., Smith, C., Thu, C.A., Tapia, J.C., Tasic, B., et al. (2012). Functional significance of isoform diversification in the protocadherin gamma gene cluster. Neuron *75*, 402–409.

Han, M.H., Lin, C., Meng, S., and Wang, X. (2010). Proteomics analysis reveals overlapping functions of clustered protocadherins. Mol. Cell. Proteomics *9*, 71–83.

Harrison, O.J., Jin, X., Hong, S., Bahna, F., Ahlsen, G., Brasch, J., Wu, Y., Vendome, J., Felsovalyi, K., Hampton, C.M., et al. (2011). The extracellular architecture of adherens junctions revealed by crystal structures of type I cadherins. Structure *19*, 244–256.

Hattori, D., Millard, S.S., Wojtowicz, W.M., and Zipursky, S.L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. Annu. Rev. Cell Dev. Biol. *24*, 597–620.

Hattori, D., Chen, Y., Matthews, B.J., Salwinski, L., Sabatti, C., Grueber, W.B., and Zipursky, S.L. (2009). Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms. Nature *461*, 644–648.

Hughes, M.E., Bortnick, R., Tsubouchi, A., Bäumer, P., Kondo, M., Uemura, T., and Schmucker, D. (2007). Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron *54*, 417–427.

Jin, X., Walker, M.A., Felsövályi, K., Vendome, J., Bahna, F., Mannepalli, S., Cosmanescu, F., Ahlsen, G., Honig, B., and Shapiro, L. (2012). Crystal structures of Drosophila N-cadherin ectodomain regions reveal a widely used class of Ca^2+-free interdomain linkers. Proc. Natl. Acad. Sci. USA *109*, E127–E134.

Lefebvre, J.L., Kostadinov, D., Chen, W.V., Maniatis, T., and Sanes, J.R. (2012). Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature *488*, 517–521.

Lommel, M., Winterhalter, P.R., Willer, T., Dahlhoff, M., Schneider, M.R., Bartels, M.F., Renner-Müller, I., Ruppert, T., Wolf, E., and Strahl, S. (2013). Protein O-mannosylation is crucial for E-cadherin-mediated cell adhesion. Proc. Natl. Acad. Sci. USA *110*, 21024–21029.

Matthews, B.J., Kim, M.E., Flanagan, J.J., Hattori, D., Clemens, J.C., Zipursky, S.L., and Grueber, W.B. (2007). Dendrite self-avoidance is controlled by Dscam. Cell *129*, 593–604.

Miura, S.K., Martins, A., Zhang, K.X., Graveley, B.R., and Zipursky, S.L. (2013). Probabilistic splicing of Dscam1 establishes identity at the level of single neurons. Cell *155*, 1166–1177.

Morishita, H., Umitsu, M., Murata, Y., Shibata, N., Udaka, K., Higuchi, Y., Akutsu, H., Yamaguchi, T., Yagi, T., and Ikegami, T. (2006). Structure of the cadherin-related neuronal receptor/protocadherin-alpha first extracellular cadherin domain reveals diversity across cadherin families. J. Biol. Chem. *281*, 33650–33663.

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics *21*, 1472–1478.

Posy, S., Shapiro, L., and Honig, B. (2008). Sequence and structural determinants of strand swapping in cadherin domains: do all cadherins bind through the same adhesive interface? J. Mol. Biol. *378*, 954–968.

Reiss, K., Maretzky, T., Haas, I.G., Schulte, M., Ludwig, A., Frank, M., and Saftig, P. (2006). Regulated ADAM10-dependent ectodomain shedding of gamma-protocadherin C3 modulates cell-cell adhesion. J. Biol. Chem. *281*, 21735–21744.

Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in the protocadherin-alpha gene cluster. Proc. Natl. Acad. Sci. USA *103*, 19719–19724.

Schalm, S.S., Ballif, B.A., Buchanan, S.M., Phillips, G.R., and Maniatis, T. (2010). Phosphorylation of protocadherin proteins by the receptor tyrosine kinase Ret. Proc. Natl. Acad. Sci. USA *107*, 13894–13899.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671–684.

Schreiner, D., and Weiner, J.A. (2010). Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proc. Natl. Acad. Sci. USA *107*, 14893–14898.

Schwartz, J.C., Zhang, X., Fedorov, A.A., Nathenson, S.G., and Almo, S.C. (2001). Structural basis for co-stimulation by the human CTLA-4/B7-2 complex. Nature *410*, 604–608.

Soba, P., Zhu, S., Emoto, K., Younger, S., Yang, S.J., Yu, H.H., Lee, T., Jan, L.Y., and Jan, Y.N. (2007). Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. Neuron *54*, 403–416.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol. Cell *10*, 21–33.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between α, β, and γ protocadherins. Cell *158*, 1045–1059.

Vester-Christensen, M.B., Halim, A., Joshi, H.J., Steentoft, C., Bennett, E.P., Levery, S.B., Vakhrushev, S.Y., and Clausen, H. (2013). Mining the O-mannose glycoproteome reveals cadherins as major O-mannosylated glycoproteins. Proc. Natl. Acad. Sci. USA *110*, 21018–21023.

Wojtowicz, W.M., Wu, W., Andre, I., Qian, B., Baker, D., and Zipursky, S.L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell *130*, 1134–1145.

Wu, Q., and Maniatis, T. (1999). A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell *97*, 779–790.

Wu, Y., Vendome, J., Shapiro, L., Ben-Shaul, A., and Honig, B. (2011). Transforming binding affinities from three dimensions to two with application to cadherin clustering. Nature *475*, 510–513.

Wu, Y., Honig, B., and Ben-Shaul, A. (2013). Theory and simulations of adhesion receptor dimerization on membrane surfaces. Biophys. J. *104*, 1221–1229.

Zipursky, S.L., and Grueber, W.B. (2013). The molecular basis of self-avoidance. Annu. Rev. Neurosci. *36*, 547–568.

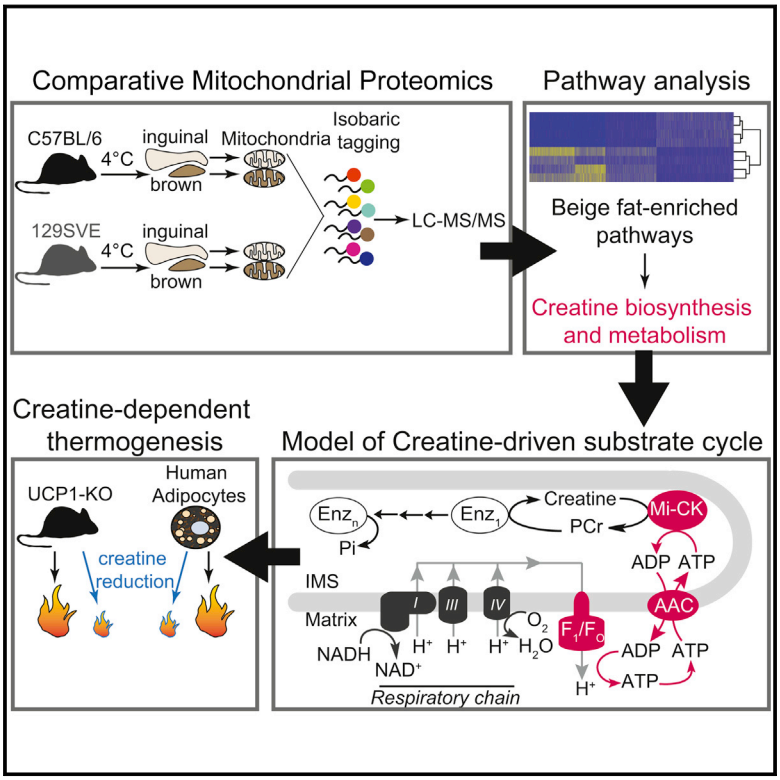
Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell *143*, 343–353.

Article

Cell

A Creatine-Driven Substrate Cycle Enhances Energy Expenditure and Thermogenesis in Beige Fat

Graphical Abstract



Highlights

- Quantitative proteomics identifies a creatine enzyme signature in beige fat
- Creatine-driven substrate cycling enhances beige-fat mitochondrial respiration
- Genes and proteins of creatine metabolism exhibit a reciprocal relationship with Ucp1
- Creatine reduction decreases energy expenditure in mice and human brown adipocytes

Authors

Lawrence Kazak, Edward T. Chouchani, Mark P. Jedrychowski, ..., Shingo Kajimura, Steve P. Gygi, Bruce M. Spiegelman

Correspondence

bruce_spiegelman@dfci.harvard.edu

In Brief

Beige fat uses creatine to dissipate energy and stimulate mitochondrial ATP demand, thereby promoting cold adaptation.

Kazak et al., 2015, Cell *163*, 643–655
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.035

CellPress

Article

Cell

A Creatine-Driven Substrate Cycle Enhances Energy Expenditure and Thermogenesis in Beige Fat

Lawrence Kazak,[1,2] Edward T. Chouchani,[1,2] Mark P. Jedrychowski,[2] Brian K. Erickson,[2] Kosaku Shinoda,[3] Paul Cohen,[1,2] Ramalingam Vetrivelan,[4] Gina Z. Lu,[1] Dina Laznik-Bogoslavski,[1] Sebastian C. Hasenfuss,[1,2] Shingo Kajimura,[3] Steve P. Gygi,[2] and Bruce M. Spiegelman[1,2,*]

[1]Dana-Farber Cancer Institute, Boston, MA 02115, USA

[2]Department of Cell Biology, Harvard University Medical School, Boston, MA 02115, USA

[3]Diabetes Center, University of California, San Francisco (UCSF), San Francisco, CA 94143, USA

[4]Department of Neurology, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

*Correspondence: bruce_spiegelman@dfci.harvard.edu

http://dx.doi.org/10.1016/j.cell.2015.09.035

SUMMARY

Thermogenic brown and beige adipose tissues dissipate chemical energy as heat, and their thermogenic activities can combat obesity and diabetes. Herein the functional adaptations to cold of brown and beige adipose depots are examined using quantitative mitochondrial proteomics. We identify arginine/creatine metabolism as a beige adipose signature and demonstrate that creatine enhances respiration in beige-fat mitochondria when ADP is limiting. In murine beige fat, cold exposure stimulates mitochondrial creatine kinase activity and induces coordinated expression of genes associated with creatine metabolism. Pharmacological reduction of creatine levels decreases whole-body energy expenditure after administration of a β3-agonist and reduces beige and brown adipose metabolic rate. Genes of creatine metabolism are compensatorily induced when UCP1-dependent thermogenesis is ablated, and creatine reduction in *Ucp1*-deficient mice reduces core body temperature. These findings link a futile cycle of creatine metabolism to adipose tissue energy expenditure and thermal homeostasis.

INTRODUCTION

Non-shivering thermogenesis primarily takes place in brown and beige adipose tissues. The ability of these depots to dissipate chemical energy has led to interest in their ability to combat obesity and diabetes. The thermogenic property of brown and beige fat relies predominantly on the actions of uncoupling protein 1 (UCP1) (Cannon and Nedergaard, 2004). This protein resides in the mitochondrial inner membrane and stimulates thermogenesis by dissipating the protonmotive force (Δp) and increasing the rate of substrate flux through the mitochondrial respiratory chain.

It is now appreciated that there are at least two distinct UCP1-expressing cell types. Classical brown adipocytes are derived from a *Myf5*$^+$ lineage and are located primarily in developmentally formed depots in the interscapular region of rodents and human infants (Seale et al., 2008). Beige-fat cells arise primarily from a *Myf5*$^-$ lineage and generally accumulate in white fat depots upon cold challenge or with the application of a number of different adrenergic stimuli, hormones, and peptide factors. Much (but probably not all) of the "brown fat" of adult humans in the neck and supraclavicular areas has the molecular characteristics of beige fat, rather than those of the classical brown fat of rodents (Shinoda et al., 2015; Wu et al., 2012). On the other hand, the interscapular depots of human infants do indeed resemble the classical brown fat of rodents (Lidell et al., 2013).

Ablation experiments in mice have shown that UCP1$^+$ cells, taken as a whole, protect mice from the metabolic effects of high-fat feeding. Diminution of UCP1$^+$ cells via transgenic expression of a toxigene first established the key role of these cells in the regulation of metabolic health (Lowell et al., 1993). Mice with *Ucp1* deletion also develop obesity, although in this case obesity is only observed at thermoneutrality (Feldmann et al., 2009). Recently, beige-fat function has been ablated in mice, with classical brown-fat function left largely intact (Cohen et al., 2014); these animals develop moderate obesity and insulin resistance centered on the liver.

The realization that mammals have two distinct thermogenic cell types raises questions regarding their similarities and their differences. Questions concerning fuel preferences, hormone sensitivities, and other key thermogenic pathways and functions are largely unexplored. We have performed quantitative proteomics, comparing highly purified mitochondria from brown and beige-fat depots. The results indicate that beige-fat cells have a thermogenic mechanism built around a creatine-driven substrate cycle.

RESULTS

Mitochondrial Purification from Cold-Exposed Brown and Beige Fat

We set out to compare the proteomic and bioenergetic properties of mitochondria isolated from beige and brown adipose tissue upon induction of thermogenesis through cold exposure. To this end, we exposed mice to 4°C, which is sufficient to drive thermogenesis in subcutaneous inguinal white adipose tissue (iWAT) and classical interscapular brown adipose tissue (BAT).

We used western blotting to evaluate the purity of our mitochondrial preparations. As expected, mitochondrial proteins were enriched, whereas contaminating components of the cytoplasm and endoplasmic reticulum were largely removed during the purification procedure (Figures S1A and S1B). Mitochondrial yield increased substantially (beige: 10-fold and brown: 2-fold) following cold exposure (Figure S1C). These mitochondria from both sources displayed properties indicative of UCP1^{+} organelles, including the requirement for BSA and purine nucleotides to acquire respiratory control (Figure S1D). These data are in line with previous reports (Shabalina et al., 2013) and indicate that beige-fat mitochondria from iWAT are functionally thermogenic following cold exposure.

Quantitative Mitochondrial Proteomics Identifies Arginine/Creatine Metabolism as a Signature of Beige Adipose Tissue

We used tandem mass spectrometry after isobaric peptide tagging to identify protein species exhibiting differential abundance between beige and brown adipose mitochondria from two strains (C57BL/6 and 129SVE) of cold-exposed mice (Figure 1A and Table S1). As shown by the global heatmap (Figure 1B) and the principal-component analysis (Figure 1C), brown-fat mitochondria showed marginal strain-dependent variance, whereas beige-fat mitochondria demonstrated greater diversity across strains. Next, beige and brown-fat mitochondrial proteomes were stratified according to differential relative abundance (Figure 1B), followed by identification of beige-fat-selective biological pathways (Figure 1D). Because mitochondrial proteins made up a larger percentage of the material after isolating pure organelles through a sucrose gradient, relative to crude mitochondria obtained by differential centrifugation alone (Figures S1A and S1B), we defined bona fide mitochondrial proteins to be those with higher abundance in the pure fraction, relative to the crude fraction. Thus, protein abundance between crude and pure preparations of mitochondria from beige and brown fat was examined by mass spectrometry (Figure 1E and Table S2). Proteins that had higher abundance in pure mitochondria, relative to crude mitochondria, were cross-referenced to the initial proteomics inventory (Table S1). Pathway analysis demonstrated that components of arginine-dependent creatine and proline metabolism were reproducibly enriched in beige-fat relative to brown-fat mitochondria (Figures 1D and 1F). A total of 14 proteins were identified that could be assigned to this pathway (Table S3). Enzymes with the ability to synthesize creatine and remove ornithine, an inhibitor creatine biosynthesis (Sipilä, 1980), showed beige-fat selectivity (Figure 1G). There was a strong correlation between western blotting and proteomic quantification for beige- and brown-enriched mitochondrial proteins (Figures S1E–S1G). Increased protein abundance in beige-fat mitochondria was observed for components of arginine-dependent creatine and proline metabolism, such as GATM and CKMT2, as well as the majority of ATP synthase subunits (Figure 1H and Table S1). Beige-fat mitochondria contained higher levels of creatine kinase (CK) activity relative to brown-fat mitochondria in both C57BL/6 and 129SVE strains of mice following cold exposure (Figure S1H). Moreover, mitochondrial CK activity was cold inducible (~2-fold) in beige fat (Figure 1I).

Given that creatine metabolism was found to be a distinct feature of thermogenic beige adipocytes at the protein level, we monitored changes in the mRNA expression of genes involved in creatine metabolism following 6 hr and 1 week exposure to 4°C. Transcript levels of these genes were coordinately elevated in response to cold in iWAT but not in BAT (Figure 1J). *Gatm* and *Ckmt1* transcript abundance was similar between iWAT and BAT. However, GATM and CKMT1 protein expression was higher in beige-fat than in brown-fat mitochondria (Table S3). In contrast, *Ckmt2* transcript levels were greater in BAT compared to iWAT. However, the expression level of CKMT2 protein was found to be slightly greater in beige-fat than in brown-fat mitochondria (Figure 1H and Table S3). The discordance between *Ckmt2* mRNA from whole-tissue lysates and protein abundance from isolated mitochondria is likely due to higher mitochondrial content in BAT than iWAT.

We next investigated the levels of creatine and phosphocreatine (PCr) in iWAT and BAT from mice housed at 30°C or 4°C. Creatine levels in iWAT were elevated 2-fold following cold exposure (Figure S1I). In contrast, although higher steady-state creatine levels were observed in BAT, cold exposure had no detectable effect (Figure S1I). There was no difference in PCr levels in either iWAT or BAT in response to cold (Figure S1J), although a modest trend toward lower PCr levels in BAT was observed; these observations are in line with a recent report (Grimpo et al., 2014). As a consequence of these measurements, it is clear that the PCr/creatine ratio in iWAT was reduced significantly in cold-exposed animals (Figure 1K), suggesting increased creatine metabolism in beige fat. These changes in creatine levels were not observed in skeletal muscle of the same animals (Figure S1K).

Creatine Stimulates Respiration in Beige-Fat Mitochondria when ADP Is Limiting

Based on our identification of creatine metabolism as a signature of beige-fat mitochondria and due to the functional coupling of mitochondrial CK (Mi-CK) to oxidative phosphorylation through the ATP/ADP carrier (AAC) (Jacobus and Lehninger, 1973; Wyss and Kaddurah-Daouk, 2000), we posited that creatine could dissipate the mitochondrial ATP pool to drive ADP-dependent respiration in beige-fat mitochondria. Such a pathway would require creatine and CK-mediated hydrolysis of ATP to drive a catalytic mechanism that stimulates cycling of ATP production and consumption (Figure 2A). We therefore tested whether creatine could stimulate substrate cycling and increase ADP-dependent respiration in beige-fat mitochondria.

Striated muscle tissues are understood to utilize creatine metabolism such that mitochondrial ATP and creatine generate PCr and ADP in a 1:1 stoichiometry (Wyss and Kaddurah-Daouk, 2000). The resulting PCr pool is used to drive substrate-level phosphorylation of ADP during times of ATP deficit. Thus, a direct prediction of this classical metabolic utilization of PCr is that addition of a quantity of creatine to oxidatively coupled mitochondria will result in a molar equivalent production of ADP and PCr through CK-mediated phosphotransferase activity (Jacobus and Lehninger, 1973). Alternatively, if creatine drives futile substrate cycling, addition of a given

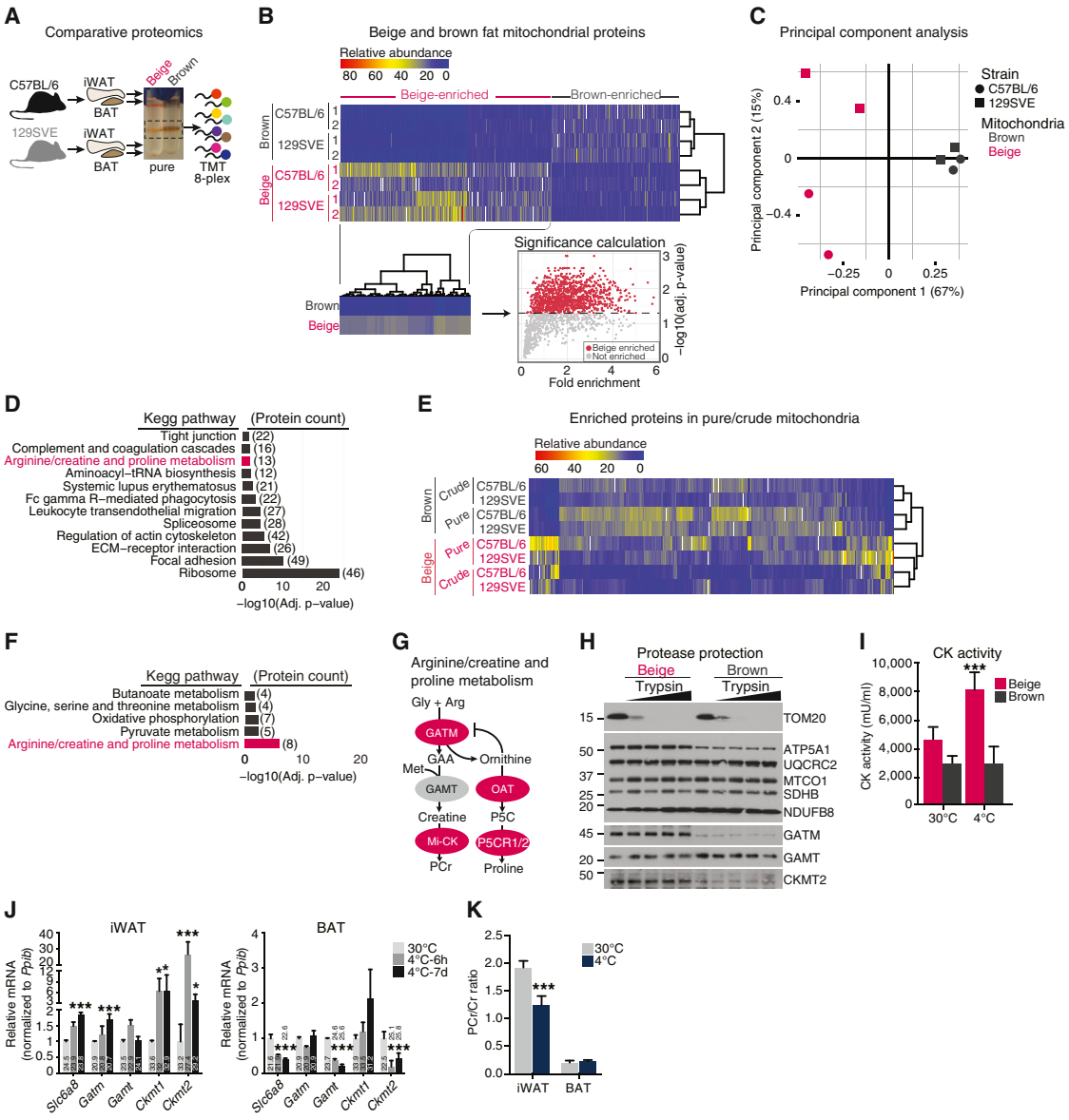


Figure 1. Characterization of Mitochondria from Brown and Beige Adipose Tissues

(A) Schematic of mitochondrial purification and quantitative proteomics workflow.

(B) Heatmap of beige-fat and brown-fat mitochondrial proteomics data (from Table S1). Beige-enriched proteins are shown in the subset heatmap.

(C) Principal-component analysis of the mitochondrial proteomics dataset.

(D) Kegg pathway analysis of beige-fat-selective mitochondrial proteins from Figure 1B.

(E) Heatmap of beige-fat and brown-fat mitochondrial proteomics data (from Table S2) after selecting proteins on the basis of an expression ratio greater than 1 in pure/crude mitochondria.

(F) Kegg pathway analysis of significantly enriched beige-fat mitochondrial proteins after cross-referencing Tables S1 and S2.

(G) Schematic of creatine synthesis and byproduct removal proteins identified by mass spectrometry. Red circles, proteins identified by mass spectrometry; gray circles, proteins not identified. Gly, glycine; Arg, arginine; Met, methionine; PCr, phosphocreatine; P5C, 1-pyrroline-5-carboxylic acid; Mi-CK, mitochondrial CK.

(H) Western blot after treatment of beige- and brown-fat mitochondria with trypsin (0, 10, 25, 50, and 100 μg ml^{-1}).

(I) CK activity of mitochondria from 129SVE mice housed at 30°C or 4°C for 7 days.

(J) Quantitative RT-PCR (qRT-PCR) from C57BL/6 mice housed at 30°C or 4°C for 6 hr (4°C—6 hr) or 7 days (4°C—7 days); n = 3 to 4 mice per group.

(K) PCr to creatine (Cr) ratio in iWAT and BAT from 129SVE mice housed at 30°C or 4°C for 7 days.

Data are presented as means ± SEM. *p < 0.05, ***p < 0.01.

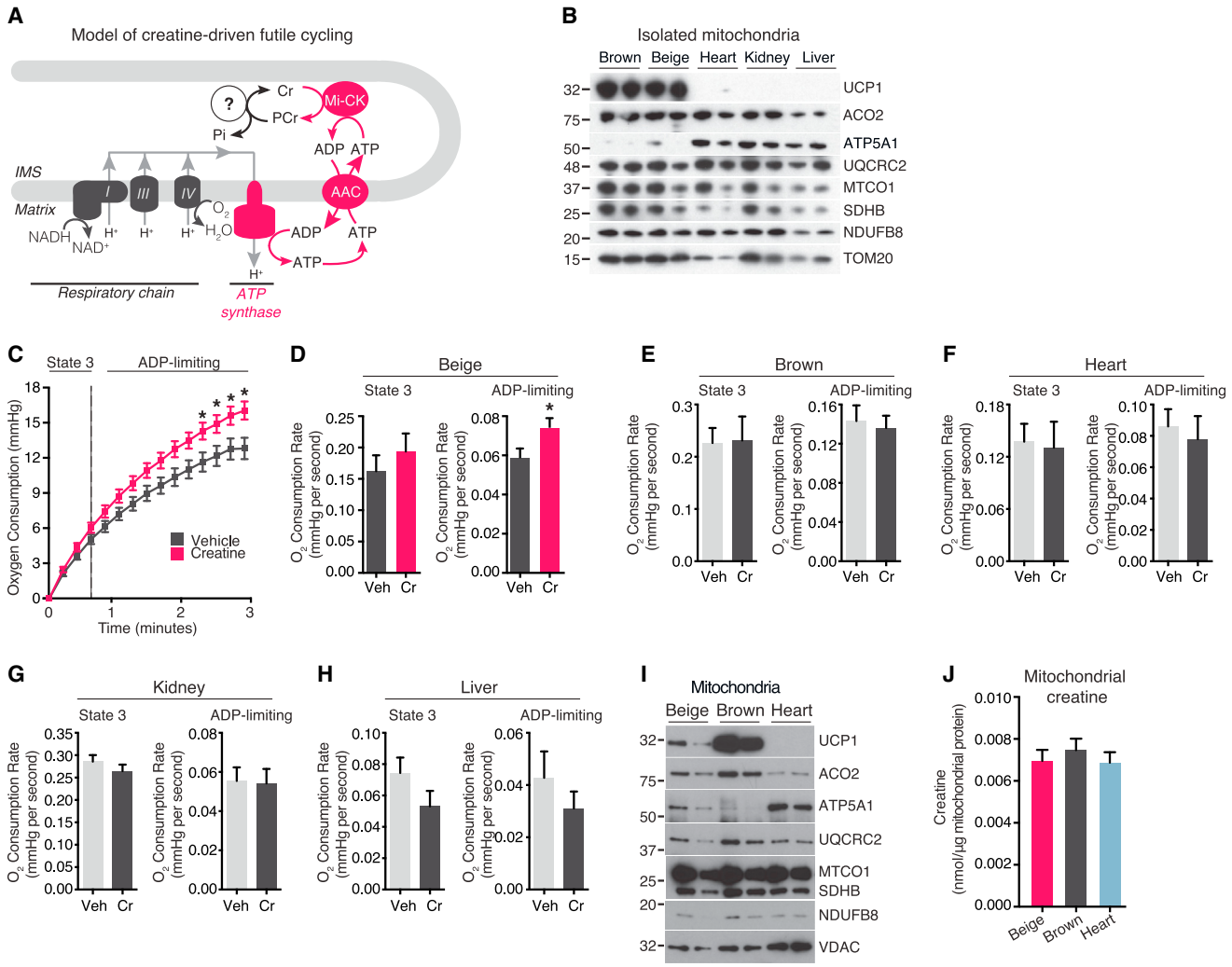


Figure 2. Creatine Stimulates Respiration in Beige-Fat Mitochondria when ADP Is Limiting

(A) Model of creatine-based substrate cycling. IMS, intermembrane space; AAC, ADP/ATP carrier.

(B) Western blot of mitochondrial proteins. n = 2 mitochondrial preparations, 15 mice per cohort.

(C) Oxygen consumption by beige-fat mitochondria treated with and without creatine (0.01 mM) in the presence of 0.2 mM ADP. Vertical dashed line, state 3 to state 4 transition. n = 9 mitochondrial preparations, 15 mice per cohort.

(D) State 3 and ADP-limiting oxygen consumption rate (OCR) of beige-fat mitochondria treated with and without creatine in the presence of 0.2 mM ADP. n = 9 mitochondrial preparations, 15 mice per cohort.

(E) State 3 and ADP-limiting OCR of brown-fat mitochondria treated as in (D). n = 3 mitochondrial preparations, 15 mice per cohort.

(F) State 3 and ADP-limiting OCR of heart mitochondria treated as in (D). n = 3 mitochondrial preparations, 8 mice per cohort.

(G) State 3 and ADP-limiting OCR of kidney mitochondria treated as in (D). n = 3 mitochondrial preparations, 15 mice per cohort.

(H) State 3 and ADP-limiting OCR of liver mitochondria treated as in (D). n = 3 mitochondrial preparations, 2 mice per cohort.

(I) Western blot of mitochondrial proteins from beige fat, brown fat, and heart. n = 2 mitochondrial preparations, 15 mice per cohort.

(J) Mitochondrial creatine concentration in beige fat, brown fat, and heart.

Data are presented as means ± SEM. *p < 0.05.

amount of creatine would instead drive the release of a molar excess of ADP with respect to creatine, thus resulting in a surplus of oxygen consumption if ADP is limiting. In this case, the relationship of creatine to ADP liberation would be substoichiometric.

The stoichiometric relationship between creatine and ATP synthase-coupled respiration was examined in mitochondria isolated from a panel of tissues. Western blotting demonstrated similar mitochondrial yields from brown fat, beige fat, heart, kidney, and liver (Figure 2B). Mitochondria were respired on pyruvate and malate in the presence of varying ADP concentrations. These organelles exhibited the expected behavior on addition of sub-saturating amounts of ADP (0.01–0.2 mM), such that respiration transitioned to state 4 as ADP became limiting (Figures S2A and S2B). In contrast, a saturating amount of ADP (1 mM) resulted in state 3 respiration with no transition to state 4, as

ADP was no longer limiting over the course of the incubation (Figure S2A).

In the presence of sub-saturating levels of ADP, addition of a sub-stoichiometric amount of creatine (0.01 mM) stimulated an ~30% increase in respiration in beige-fat mitochondria. Specifically, creatine did not significantly affect the initial (state 3) ADP-stimulated rate of oxygen consumption (Figures 2C and 2D, left panel) but enhanced the respiration rate when ADP levels became limiting (Figures 2C and 2D, right panel). This action of creatine in beige-fat mitochondria did not occur in the absence of exogenous ADP (Figure S2C) nor did it occur in the presence of the AAC inhibitor, carboxyatractyloside (Figure S2D). These data suggest that creatine-mediated respiration required export of mitochondrial matrix ATP through the AAC upon addition of exogenous ADP (Figure 2A). This feature of creatine-dependent respiration is in agreement with the established functional coupling of AAC with Mi-CK isoforms (Jacobus and Lehninger, 1973; Wyss and Kaddurah-Daouk, 2000). Furthermore, incubation of beige-fat mitochondria with saturating ADP concentrations (Figure S2E) precluded the respiration-stimulating effects of creatine (Figure S2F), further confirming that the stimulatory effects occurred only under ADP-limiting conditions.

On the basis of the mitochondrial P/O ratio (Watt et al., 2010) we calculated the amount of additional ATP that was synthesized by addition of creatine to beige-fat mitochondria. Creatine drove an increase in oxygen consumption equivalent to a 7.82-fold (±2.12) excess phosphorylation of ADP. These results indicate that in beige adipose mitochondria, creatine acts substoichiometrically (with respect to ADP phosphorylation) to increase mitochondrial ATP synthesis and stimulate substrate flux through the mitochondrial respiratory chain. Addition of sub-stoichiometric amounts of creatine to mitochondria isolated from brown fat, heart, kidney, or liver (Figure 2B) had no effect on ADP-dependent respiration (Figures S2G and 2E–2H).

Striated muscle contains a large quantity of creatine (Fitch and Chevli, 1980), and so endogenous mitochondrial creatine levels in beige and brown fat were compared to heart following purification of a comparable yield of organelles from these tissues (Figure 2I). Figure 2J demonstrates that mitochondrial creatine levels were similar between tissues following purification, indicating that the lack of a respiration-enhancing effect in heart mitochondria was not due to a high amount of endogenous creatine. Together, these data indicate that creatine stimulates respiration in beige-fat mitochondria in a substoichiometric manner with respect to ADP, specifically when ADP is limiting; this is consistent with a model of creatine-driven substrate cycling (Figure 2A).

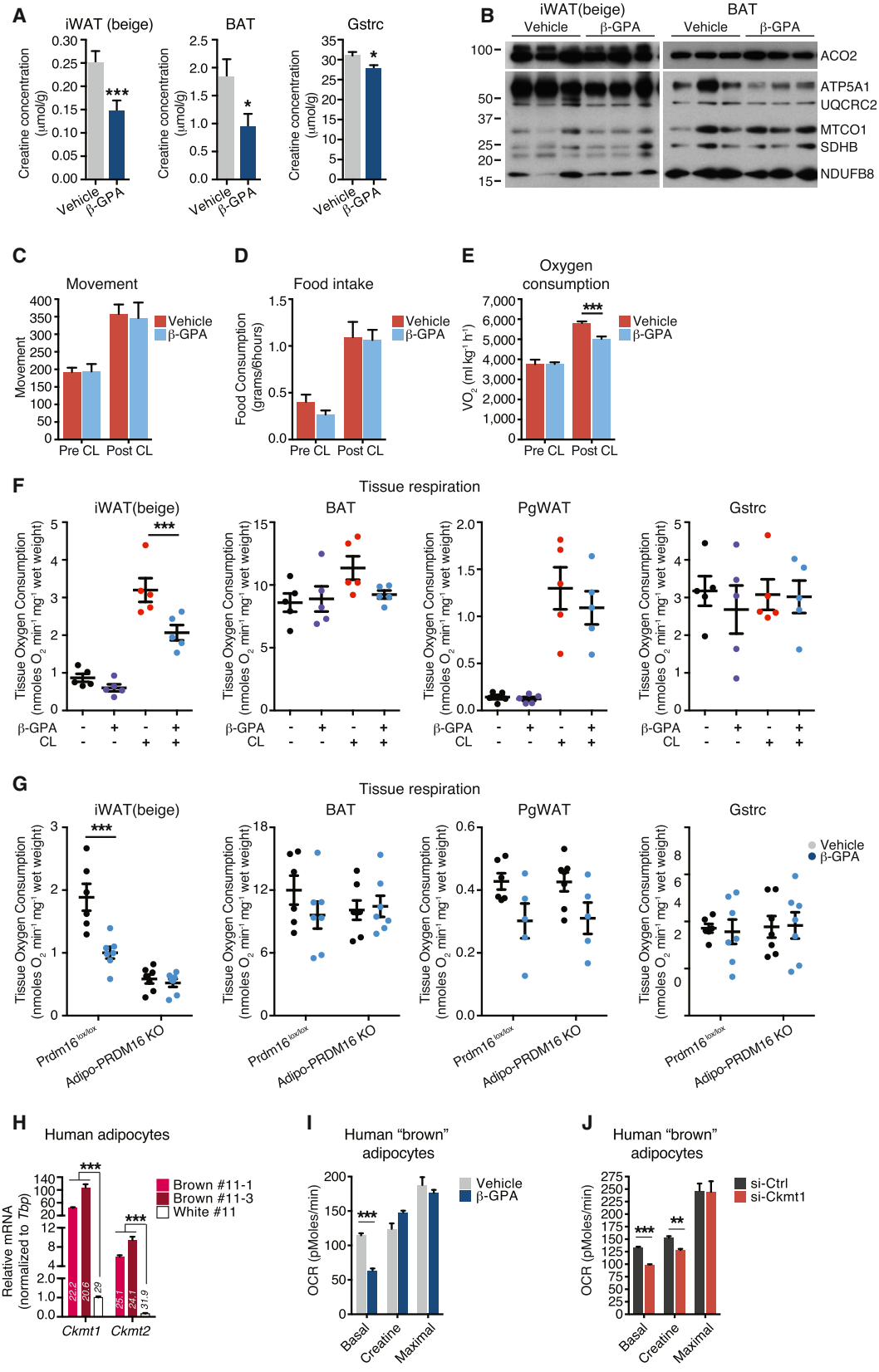
To examine thermogenesis by creatine-driven substrate cycling, we employed differential scanning calorimetry (Ricquier et al., 1979). As expected, chemical uncoupling by FCCP induced thermogenesis, and this signal was inhibited by rotenone and antimycin; the signal was similar to that obtained when mitochondria were omitted from the reaction (Figure S2H). Importantly, addition of creatine to beige-fat mitochondria drove ADP-dependent thermogenesis by ~30% relative to organelles incubated with ADP alone (Figure S2I). This is a direct demonstration of thermogenesis through creatine metabolism in beige-fat mitochondria.

Creatine Metabolism in Adipose Tissue Contributes to Energy Expenditure and Thermal Homeostasis In Vivo

As the role of creatine metabolism in thermogenic fat tissues is essentially unexplored in vivo, we systematically assessed its role in beige and brown adipose metabolic functions. First, we examined whether creatine regulates oxidative metabolism in beige and brown fat. To this end, we utilized β-guanidinopropionic acid (β-GPA), a creatine analog that is well established to inhibit creatine transport and to reduce creatine levels in cultured cells and tissues (Fitch and Chevli, 1980). A diet supplemented with a typical dose of β-GPA (2%), resulted in reductions in food intake and weight loss, as previously reported (Oudman et al., 2013); this compromised subsequent metabolic analyses. However, when mice were given daily intraperitoneal injections with a lowered dose of β-GPA (0.4 g kg^{-1}) for 4 days during cold exposure, body weight and fat-pad weight were unaffected (Figures S3A and S3B). Creatine levels were reduced by approximately 50% in iWAT and BAT and by 15% in gastrocnemius muscle (Figure 3A). Mitochondrial respiratory chain protein expression was not altered (Figure 3B). To determine the contribution of creatine metabolism in brown/beige adipose to whole-body energy expenditure, we utilized the β3-adrenergic receptor agonist CL 316,243 (CL), a well-known activator of adipose thermogenesis (Bloom et al., 1992; Granneman et al., 2003). Movement, food intake, and oxygen consumption were monitored in CL-treated mice that were co-treated with either vehicle or β-GPA. There was no difference in ambulatory movement, food intake, or fuel utilization between vehicle- or β-GPA-treated animals (Figures 3C, 3D, and S3C). As expected, we detected a 54% increase in the metabolic rate of mice following CL injection. Strikingly, a reduction in creatine through the administration of β-GPA diminished the CL-induced increase in whole-body oxygen consumption by ~40% compared to vehicle treatment (Figure 3E).

The reduction in whole-body oxygen consumption prompted an examination of which tissues were most affected by β-GPA, in terms of whole-tissue respiration. Thus, another cohort of mice, housed at 23°C, was separated into four groups: vehicle, β-GPA, CL, or CL + β-GPA. β-GPA had no effect on the respiration of any tissue examined in the absence of a browning stimulus (Figure 3F). CL treatment provided a powerful browning/beiging stimulus, resulting in a 5-fold increase in iWAT oxygen consumption, a 33% increase in BAT respiration, a 9-fold elevation in PgWAT respiration, and no effect on Gstrc respiration (Figure 3F). Reducing creatine levels with β-GPA had a profound and significant effect on beige iWAT, resulting in a 34% reduction in oxygen consumption (Figure 3F). BAT oxygen consumption was reduced by 18% following β-GPA treatment, PgWAT oxygen consumption was decreased by 15%, and no detectable effect on Gstrc tissue respiration was detected (Figure 3F). Although the β-GPA-dependent reduction in BAT respiration did not reach statistical significance, the absolute decrease was larger than any other tissue examined, and so in addition to iWAT, BAT may well contribute to the effects of β-GPA observed in vivo.

We next examined the contribution of beige adipocytes within iWAT to creatine-dependent oxidative metabolism. Mice with an adipose-selective deletion of *Prdm16* (Adipo-PRDM16 KO) have disrupted beige adipose function (Cohen et al., 2014). Following



(legend on next page)

7 days of cold exposure, creatine reduction with β-GPA significantly decreased iWAT oxidative metabolism from control *Prdm16*$^{lox/lox}$ mice but had no further effect on the already reduced respiratory rate of iWAT from Adipo-PRDM16 KO mice (Figure 3G). No significant effect of creatine reduction was detected in any other tissue in either genotype, although again, there was a modest reduction in the respiration of BAT from *Prdm16*$^{lox/lox}$ mice (Figure 3G). Therefore, using two models of browning/beiging, these data demonstrate that creatine reduction attenuates the oxidative metabolism of thermogenic adipose tissues in vivo.

Respiration of Cultured Human Brown Adipocytes Is Regulated by Creatine

Much, though not all, of the brown fat observed in adult humans has characteristics of murine beige adipocytes (Lidell et al., 2013; Shinoda et al., 2015; Wu et al., 2012). Thus, it was of interest to examine whether creatine contributed to oxidative metabolism in the recently cloned human brown adipocytes (Shinoda et al., 2015). Both *Ckmt1* and *Ckmt2* mRNA levels were higher in two human clonal brown adipocyte lines compared to human white adipocytes (Figure 3H). When human brown adipocytes were treated with β-GPA, basal respiration was significantly reduced by 45%; this reduction was completely rescued by creatine supplementation (Figure 3I), demonstrating the specificity of β-GPA. Next, RNAi-mediated knockdown was used to assess the function of *Ckmt1* in these adipocytes. Transfection of small interfering RNAs (siRNAs) targeting *Ckmt1* resulted in an ~40% reduction in *Ckmt1* mRNA levels, relative to control siRNA-transfected cells (Figure S3D). Knockdown of *Ckmt1* resulted in a 26% reduction in basal respiration, and creatine supplementation did not rescue this effect (Figure 3J). These data demonstrate that creatine and CKMT1 regulate oxidative properties of isolated human brown adipocytes.

Compensatory Regulation of Creatine Metabolism Components and UCP1 in Adipose Tissues In Vivo

The ability of creatine to stimulate thermogenesis in an apparently sub-stoichiometric manner with respect to ADP-dependent respiration suggested that this creatine-based futile cycle could be an important mechanism of thermoregulation in vivo, independent of UCP1. To begin to test this supposition, we re-examined the longstanding and important observation that mice lacking *Ucp1* can maintain thermal homeostasis at 4°C if they are gradually acclimated to this temperature (Golozoubova et al., 2001). These data demonstrated that alternative thermogenic pathway(s) compensate for the lack of UCP1 (Ukropec et al., 2006). We therefore gradually acclimated *Ucp1*-deficient (*Ucp1*$^{-/-}$) mice to the cold over a period of 21 days. In addition, we also reduced creatine levels by administration of β-GPA to both *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ animals. As shown previously (Ukropec et al., 2006), cold-exposed *Ucp1*$^{-/-}$ mice had elevated expression of classical thermogenic genes, such as *Dio2* and *Pgc-1α*, in iWAT compared to *Ucp1*$^{+/+}$ mice (Figure 4A). *Pgc-1α* was also induced in BAT from *Ucp1*$^{-/-}$ mice (Figure S4A). These data suggest a strong induction of a thermogenic phenotype in the iWAT of *Ucp1*$^{-/-}$ mice. *Slc6a8* and *Ckmt1* transcripts were elevated in iWAT of cold-exposed *Ucp1*$^{-/-}$ mice, relative to *Ucp1*$^{+/+}$ littermates (Figure 4B), whereas *Slc6a8* and *Ckmt2* transcript levels were higher in BAT from the same animals (Figure S4B).

As the induction of creatine metabolism genes was elevated in *Ucp1*$^{-/-}$ mice, classical thermogenic genes, such as *Ucp1* and *Dio2*, were significantly elevated in iWAT of cold-exposed mice after creatine reduction with β-GPA (Figure 4C). A modest increase in *Ucp1* and *Pgc-1α* mRNA was observed in BAT of cold-exposed *Ucp1*$^{+/+}$ mice following creatine reduction with β-GPA (Figure S4C). UCP1 protein levels were also elevated in response to β-GPA in iWAT of *Ucp1*$^{+/+}$ mice but did not change in the BAT of the same animals (Figure 4D).

We also examined the compensatory regulation between creatine metabolism genes and Ucp1 in human brown adipocytes. In these cell lines, *Ckmt1* was more abundant than *Ckmt2* at the mRNA level (Figure 3H). Transfection of siRNAs targeting human *Ckmt1* resulted in an ~40% reduction in *Ckmt1* mRNA levels, relative to control siRNA-transfected cells (Figure 4E). Strikingly, *Ucp1* mRNA was elevated 2.5-fold in cells transfected with *Ckmt1*-targeted siRNAs relative to control siRNA-transfected cells. Forskolin treatment of human brown-fat cells resulted in a near 8-fold induction of *Ucp1* mRNA (Figure 4E), and combining forskolin treatment with *Ckmt1* knockdown resulted in a 15-fold elevation in *Ucp1* transcript abundance above non-treated si-Ctrl cells (Figure 4E). Furthermore, si-Ckmt1-transfected cells had an elevated abundance of *Dio2* and *Pgc-1α*, relative to si-Ctrl transfected cells, following forskolin treatment (Figure 4E). Taken together, these data demonstrate

Figure 3. Creatine Metabolism in Adipose Tissue Contributes to Energy Expenditure and Thermal Homeostasis In Vivo

(A) Creatine concentration in iWAT (beige), BAT, and gastrocnemius muscle (Gstrc) from cold-exposed C57BL/6 mice treated with four daily injections of vehicle or β-GPA (0.4 g kg^{-1}); n = 5 mice per group.

(B) Western blot of iWAT (beige) and BAT from animals treated as in (A); n = 3 mice per group.

(C) Movement, n = 8 mice per group. CL (0.2 mg kg^{-1}) was co-injected intraperitonealy with vehicle (saline) or β-GPA (0.4 g kg^{-1}).

(D) Food intake, n = 8 mice per group, treated as in (C).

(E) Oxygen consumption, n = 8 mice per group, treated as in (C).

(F) OCR of minced tissues following four daily injections of vehicle, β-GPA (0.4 g kg^{-1}), CL (0.2 mg kg^{-1}), or β-GPA + CL; n = 5 mice per group.

(G) OCR of minced tissues from *Prdm16*$^{lox/lox}$ and Adipo-PRDM16 KO mice after 7 days cold exposure. β-GPA was administered on the last 4 days of cold exposure; n = 5 to 7 mice per group.

(H) qRT-PCR of *Ckmt1* and *Ckmt2* in two human brown adipocyte lines (#11-1 and 11-3) and one human white adipocyte line (#11). Raw ct values are embedded in the bars.

(I) OCR of human brown adipocytes treated with vehicle, β-GPA (2 mM), and creatine (5 mM).

(J) OCR of human brown adipocytes transfected with si-Ctrl or si-Ckmt1. Creatine used at 5 mM.

Data are presented as means ± SEM. *p < 0.05, **p < 0.01, ***p < 0.0001.

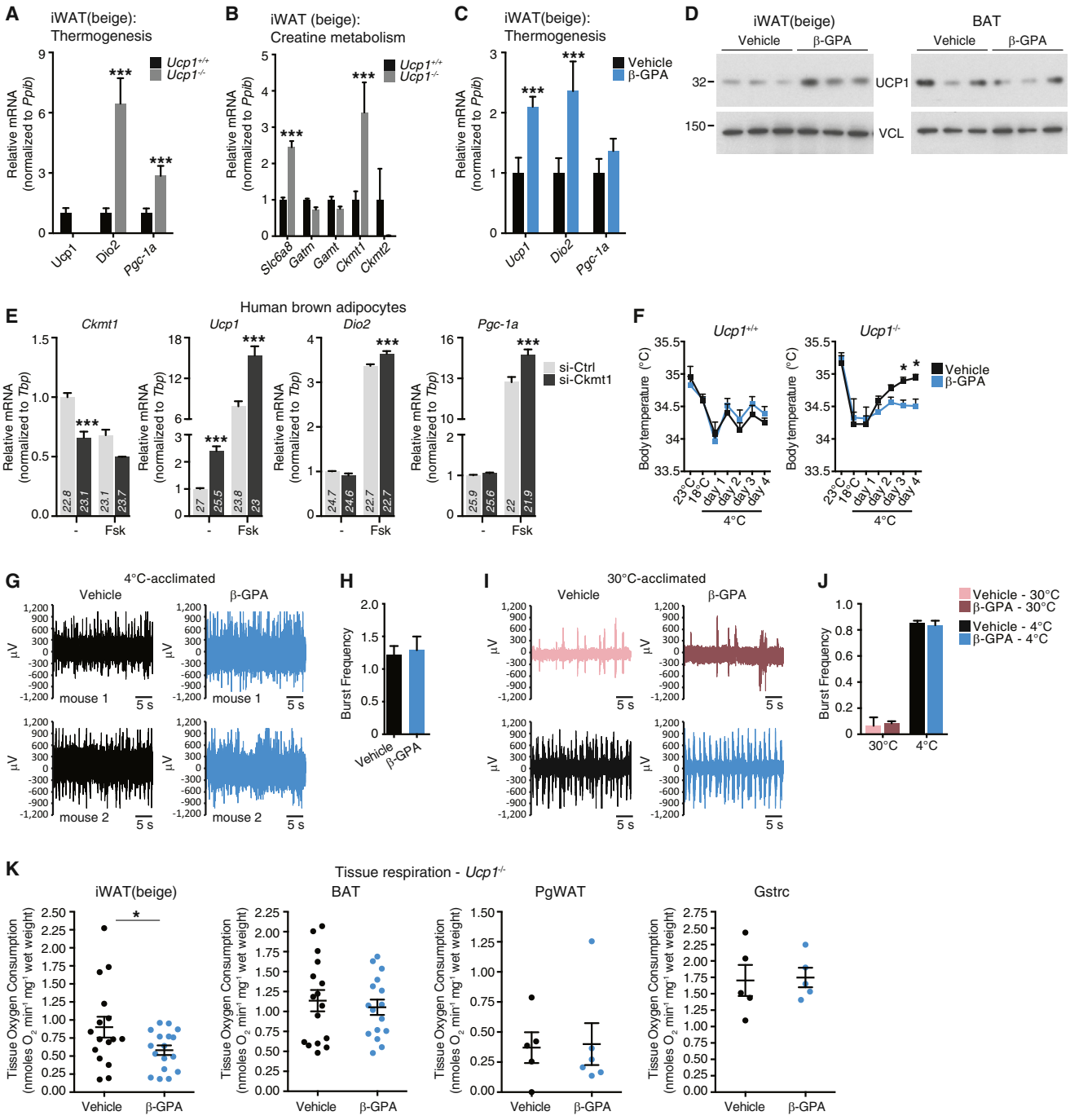


Figure 4. Creatine Metabolism Regulates UCP1-Independent Thermal Homeostasis In Vivo

(A) qRT-PCR from iWAT of 4°C-acclimated *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice; n = 5 mice per group.

(B) qRT-PCR of creatine metabolism genes from same samples as in (A).

(C) qRT-PCR from iWAT of 4°C-acclimated mice treated with four daily injections of vehicle or β-GPA (0.4 g kg^{-1}); n = 5 mice per group.

(D) Western blot from iWAT and BAT from mice treated as in (C). Vinculin (VCL), loading control, n = 3 mice per group.

(E) qRT-PCR of clonal human brown adipocytes (line #11-1). siRNAs targeted against control (si-Ctrl) or Ckmt1 (si-Ckmt1). Forskolin was used at 10 μM for 4 hr. (The data showing si-Ctrl and si-Ckmt1, without forskolin treatment, are the same data shown in Figure S3D); n = 3 per group.

(F) Body temperature of *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice treated with vehicle or β-GPA (0.4 g kg^{-1}); n = 7 to 8 mice per group.

(G) Representative electromyogram (EMG) traces, measured at 4°C, of 4°C-acclimated *Ucp1*$^{-/-}$ mice treated as in (F).

(H) Frequency of shivering bursts quantified from data in (G).

(I) Representative EMG traces, at 30°C and following 15–45 min at 4°C, of 30°C-acclimated wild-type C57BL/6 mice treated as in (F).

(legend continued on next page)

that genes involved in creatine metabolism are increased in the absence of UCP1, and expression of classical thermogenic genes is elevated when creatine metabolism is disrupted. These data suggest a compensatory relationship between UCP1- and creatine-dependent bioenergetics in mice and humans.

Creatine Regulates UCP1-Independent Thermal Homeostasis In Vivo

To test the role of creatine metabolism in adaptive thermogenesis in vivo, we examined the body temperature of cold-acclimated *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice that were treated with either β-GPA or vehicle for 4 days. *Ucp1*$^{-/-}$ mice weighed slightly less than *Ucp1*$^{+/+}$ mice (Figure S4D). Body temperature of *Ucp1*$^{-/-}$ mice was higher than that of *Ucp1*$^{+/+}$ animals at 23°C and 4°C (Figure 4F) (*Ucp1*$^{+/+}$: 34.4 ± 0.11°C and *Ucp1*$^{-/-}$: 34.7 ± 0.14°C, at 4°C). This temperature difference between genotypes is consistent with previous work (Liu et al., 2003). Strikingly, body temperature of the *Ucp1*$^{-/-}$ animals was significantly lower when endogenous creatine levels were reduced with β-GPA (Figure 4F), suggesting that in the absence of UCP1, creatine regulates a larger proportion of cold-induced thermogenesis than when UCP1 is present.

Creatine contributes to skeletal muscle metabolism, and so it was critical to determine whether the effect of β-GPA on body temperature was due to reductions in shivering thermogenesis. Thus, we used electromyography (EMG) to monitor shivering directly. Mice housed at 30°C showed minimal EMG activity. However, within 15 to 30 min of exposure to 4°C, bursts of shivering were detected (Figure S4E). Next, we used EMG on *Ucp1*$^{-/-}$ mice that had been acclimated to 4°C, and similar to previous work (Golozoubova et al., 2001), we detected robust shivering activity (Figure S4E). These measurements were confirmed as shivering bursts because they were abolished with the nicotinic acetylcholine receptor inhibitor D-tubocurarine (Figure S4E). The capacity for shivering of cold-acclimated *Ucp1*$^{-/-}$ mice was not altered by β-GPA treatment (Figures 4G and 4H). We next examined shivering thermogenesis and body temperature in wild-type mice that had been acclimated to 30°C and acutely cold challenged at 4°C following administration of vehicle or β-GPA for 4 days. Under these conditions, reduced creatine levels did not alter shivering capacity (Figures 4I and 4J) or body temperature (Figure S4F). Analysis of serum CK activity was used as an independent method to monitor shivering. Although serum CK activity was elevated in *Ucp1*$^{-/-}$ mice at 18°C, CK activity was elevated to a similar extent in *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice when ambient temperature was reduced to 10°C and 4°C (Figure S4G). Importantly, β-GPA administration had no effect on serum CK activity in either genotype. Together, these results indicate that the reduced body temperature of β-GPA-treated *Ucp1*$^{-/-}$ mice was not due to aberrant shivering, and that the metabolic and thermogenic adaptations that exist in the absence of UCP1 cannot be explained solely by shivering thermogenesis. Examination of respiration by isolated tissues indicated that despite a lower respiratory capacity of iWAT from *Ucp1*$^{-/-}$ mice, relative to *Ucp1*$^{+/+}$ mice (compare Figure 4K to Figure S4H), β-GPA treatment reduced oxygen consumption substantially in this depot from *Ucp1*$^{-/-}$ mice by 40% (Figure 4K). No significant effect on respiration due to creatine reduction was observed in any other tissue examined; however, a modest reduction was observed in BAT (Figures 4K and S4H).

A Screen of Mitochondrial Phosphatases to Identify Candidate Regulators of a Creatine-Driven Substrate Cycle

Taken together, our data indicated that creatine drives high-energy phosphate-dependent substrate cycling, which requires ADP-stimulated respiration. This led us to consider candidate molecules in adipose mitochondria that could carry out this process. To this end, we mined our mitochondrial proteomics inventory for putative phosphatases/hydrolases and identified 11 candidate phosphatases. Because several known members of creatine metabolism had elevated expression in *Ucp1*$^{-/-}$ mice (Figure 4B), we posited that the expression of a key enzyme(s) with phosphatase activity might also be elevated in the absence of UCP1. Transcript levels for all 11 of these phosphatases were measured in iWAT and BAT of cold-exposed *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice. One gene—*Phospho1* (phosphatase orphan 1)—was dramatically altered at the mRNA level, with an 8-fold increase in iWAT of *Ucp1*$^{-/-}$ animals (Figure 5A). In contrast, none of the candidates were increased in BAT at the mRNA level (Figure S5A). The consistent identification of PHOSPHO1 in our mitochondrial proteomic experiments suggested that PHOSHPO1 is at least partly associated with adipose mitochondria. Strikingly, protein expression of PHOSPHO1 was dramatically higher in the iWAT and BAT of *Ucp1*$^{-/-}$ mice, relative to *Ucp1*$^{+/+}$ mice (Figure 5B). PHOSPHO2, the closest homolog of PHOSPHO1, displayed the opposite expression pattern (Figure 5B). Proteomic quantification of PHOSPHO1 from iWAT and BAT of cold-acclimated *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice also demonstrated results similar to those of western blotting (Figure S5B). Interestingly, creatine reduction with β-GPA resulted in a small increase in *Phospho1* mRNA levels in iWAT from cold-exposed *Ucp1*$^{+/+}$ mice (Figure 5C), and a modest trend in the same direction was detected in BAT (Figure S5C). We also examined the compensatory regulation between Phospho1 and Ucp1 in human brown adipocytes. Transfection of siRNAs targeting human *Phospho1* resulted in an ~50% reduction in *Phospho1* mRNA levels, relative to control siRNA-transfected cells (Figure 5D). *Ucp1* mRNA was significantly elevated 1.5-fold in cells transfected with *Phospho1*-targeted siRNAs relative to control siRNA-transfected cells, and no change was observed for the differentiation marker *Pparg* (Figure 5D). Therefore, based on the reciprocal expression pattern between

(J) Frequency of shivering bursts quantified from data in (I).

(K) OCR of minced tissues from 4°C-acclimated *Ucp1*$^{-/-}$ mice treated as in (F). n = 16 to 17 mice per group (iWAT and BAT); n = 5 to 6 mice per group (PgWAT and Gstrc).

Data are presented as means ± SEM. *p < 0.05, ***p < 0.0001.

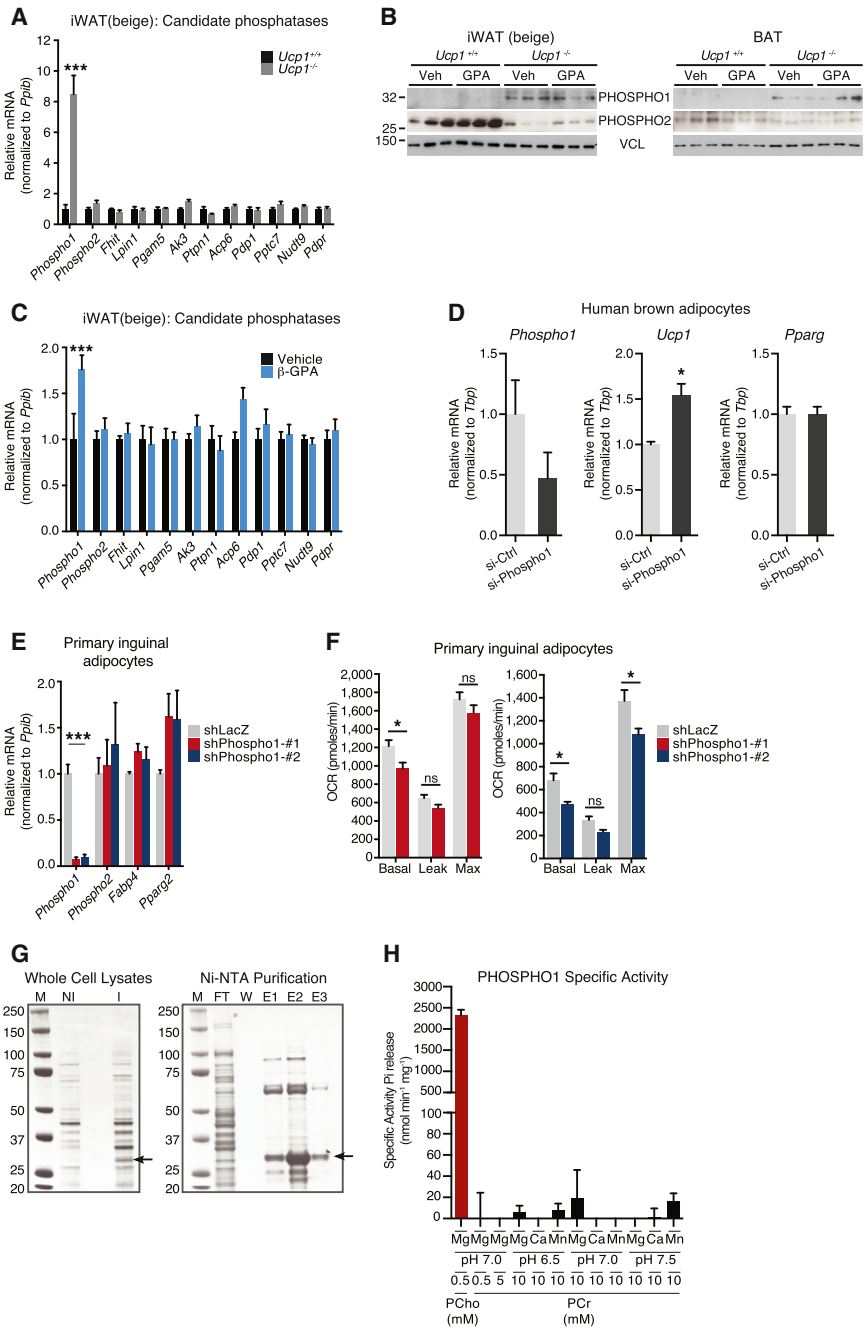


Figure 5. A Screen of Mitochondrial Phosphatases with *Ucp1*-Deficient Mice Identifies PHOSPHO1 as a Regulator of Adipocyte Respiration

(A) qRT-PCR of candidate phosphatases in iWAT of 4°C-acclimated *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice; n = 5 mice per group.

(B) Western blot of PHOSPHO1 and PHOSPHO2 from 4°C-acclimated *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice, treated with vehicle or β-GPA (0.4 g kg^{-1}). Vinculin (VCL), loading control.

(C) qRT-PCR of candidate phosphatases in iWAT of 4°C-acclimated wild-type mice treated as in (B); n = 5 mice per group.

(D) qRT-PCR of clonal human brown adipocytes (line #11-1) treated with si-Ctrl or si-Phospho1; n = 3 per group.

(E) qRT-PCR of primary mouse inguinal adipocytes after *Phospho1* knockdown (shPhospho1-#1 and shPhospho1-#2); n = 4 to 5 per group.

(F) OCR of primary mouse inguinal adipocytes treated as in (E); n = 5 to 7 per group.

(G) Coomassie stain of SDS-PAGE demonstrating PHOSPHO1 protein expression and purification. M, molecular weight marker; NI, non-induced; I, induced; FT, flow-through; W, wash; E1–E3, elutions 1–3. Arrows indicate recombinant PHOSPHO1.

(H) Specific activity of PHOSPHO1 toward PCho (0.5 mM) and PCr (0.5 to 10 mM) exposed to various buffer pH and divalent metals.

Data are presented as means ± SEM. *p < 0.05, ***p < 0.0001.

Phospho1 and *Ucp1* in mouse tissue and human cultured cells, PHOSPHO1 was considered as a candidate for involvement in a creatine-driven substrate cycle.

PHOSPHO1 Regulates Oxidative Metabolism in Primary Inguinal Adipocytes

Phospho1 displayed an expression pattern that would be predicted for a factor involved in an alternative, UCP1-independent pathway of energy expenditure. At the mRNA level, its steady-state expression was modestly higher in murine primary inguinal adipoyctes, relative to primary brown-fat cells, whereas *Ucp1* demonstrated the opposite expression pattern (Figure S5D). Therefore, the role of PHOSPHO1 in adipocyte bioenergetics was examined. Using adenoviral-mediated shRNA delivery, we achieved ~90% knockdown of *Phospho1* at the mRNA level in primary inguinal cells, with no mRNA changes detected for its homolog, *Phospho2*, or the differentiation markers *Fabp4* and *Pparg2* (Figure 5E). Reduced PHOSPHO1 levels were also observed at the protein level (Figure S5E). Both shRNAs used to knock down PHOSPHO1 significantly reduced basal oxygen consumption but had no effect on proton leak; the second shRNA also reduced maximal respiration (Figure 5F). Knockdown of *Phospho1* in primary brown adipocytes (Figure S5F) decreased basal respiration but had no effect on proton leak or maximal respiration (Figure S5G). Thus, reduced Phospho1 levels attenuated oxygen consumption in primary inguinal and brown adipocytes, although the relative effect was greater in inguinal cells. The more pronounced effect on basal respiration (largely regulated by ATP demand) than on proton leak is consistent with a role for PHOSPHO1 in

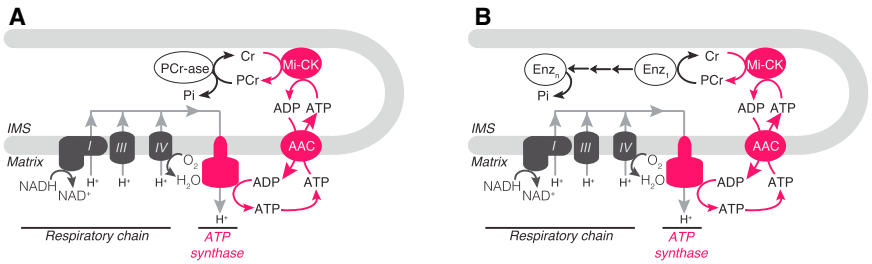


Figure 6. Models of Creatine-Driven Futile Substrate Cycling

(A) Model of creatine-driven futile substrate cycling based on direct hydrolysis of PCr.

(B) Model of creatine-driven futile substrate cycling based on multiple phosphotransfer events catalyzed by multiple enzymes (Enz_1–Enz_n).

ATP-coupled oxygen consumption. One interpretation of this data was that PHOSPHO1 regulates substrate cycling by liberating the high-energy phosphate from PCr. Therefore, we purified recombinant PHOSPHO1 (Figure 5G) in order to examine whether PCr can be directly hydrolyzed by PHOSPHO1 in vitro. PHOSPHO1-specific activity toward phosphocholine was similar to that of a prior report (Roberts et al., 2004). However, there was limited phosphatase activity toward PCr; this was done under various buffer conditions (Figure 5H). Taken together, these data suggest that if PHOSPHO1 is involved in a creatine-driven substrate cycle, it likely regulates high-energy phosphate metabolism downstream of the phosphotransfer event that utilizes PCr (Figure 6).

DISCUSSION

Although UCP1 is critical for optimal thermogenesis, *Ucp1*$^{-/-}$ animals can survive cold temperatures if gradually acclimated (Golozoubova et al., 2001; Meyer et al., 2010; Ukropec et al., 2006). Furthermore, it has been shown with adrenergic stimulation and peptide factors that the thermogenic responses of *Ucp1*$^{+/+}$ and *Ucp1*$^{-/-}$ mice are similar (Granneman et al., 2003; Grimpo et al., 2014; Véniant et al., 2015). These findings imply the presence of UCP1-independent thermogenic mechanisms. Glycerol-3-phosphate shuttle activation and lipid turnover (Flachs et al., 2011; Grimpo et al., 2014) have been posited to act independently of UCP1. Calcium cycling has been proposed to be an additional source of thermogenesis in BAT (Ukropec et al., 2006) and is a well-established thermogenic mechanism in the extraocular heater muscle cells of certain fish and in mammalian skeletal muscle (Bal et al., 2012; Block et al., 1994). Interestingly, large reductions in creatine levels have previously been linked to deregulated thermal homeostasis in rats (Wakatsuki et al., 1996; Yamashita et al., 1995), through unknown mechanisms.

The stoichiometric relationships observed between creatine, ADP, and oxygen consumption suggest a creatine-driven substrate cycle. By liberating a molar excess of ADP, with respect to the amount of added creatine, beige-fat mitochondria enhance respiration by maintaining a state of ATP synthesis, as shown by the increased rate of oxygen consumption when ADP was limiting. Reducing creatine levels by even 50% in beige and brown fat was associated with a blunted response to β3-agonism at the level of whole-body energy expenditure. A role for creatine and creatine kinase in brown adipose tissue metabolism has been posited previously (Berlet et al., 1976; Terblanche et al., 1998; Watanabe et al., 2008). Our data suggest that creatine metabolism regulates energy expenditure in both beige and brown fat.

Gene-expression analyses demonstrated a clear compensatory regulation between classical thermogenic genes and the genes involved in creatine metabolism in murine tissues and human adipocytes. Consistent with this reciprocal relationship, a reduction in creatine levels attenuated oxygen consumption of iWAT from *Ucp1*$^{-/-}$ mice and perturbed thermal homeostasis of these animals without diminishing shivering. The expression of *Phospho1* was elevated at the mRNA and protein level in *Ucp1*-deficient animals and thus became the focus of examination. Although the effect of PHOSPHO1 knockdown on adipocyte bioenergetics is consistent with a role for it in high-energy phosphate metabolism, PHOSPHO1 does not hydrolyze PCr in vitro, at least under any conditions we tested. Interestingly, phosphoethanolamine (PEtn), a direct PHOSPHO1 substrate (Roberts et al., 2004), has been demonstrated to inhibit ADP-stimulated mitochondrial respiration (Gohil et al., 2013), and so PHOSPHO1 could affect oxygen consumption by regulating PEtn levels.

The data presented here indicate that creatine metabolism plays an important role in adipose energy expenditure in vivo. Most likely, creatine facilitates the regeneration of ADP through futile hydrolysis of PCr. This could occur in a single enzymatic step, whereby PCr would be the direct substrate of a so-called PCr phosphatase (Figure 6A), or via multiple phosphotransfer reactions prior to phosphate hydrolysis from a currently unknown phosphometabolite (Figure 6B).

Similar to what was shown in the recently cloned human brown adipocytes (Shinoda et al., 2015), CKMT1 and CKMT2 expression is enriched in human BAT (Svensson et al., 2011). If creatine metabolism plays a substantial role in thermogenesis in humans, as suggested by the work here with isolated cells, it could open up possibilities to manipulate energy expenditure in patients with metabolic diseases by new drugs or even with dietary supplementation.

EXPERIMENTAL PROCEDURES

Sucrose Gradient-Purified Mitochondria

Sucrose was dissolved to a concentration of 1M and 1.5M in gradient buffer (10 mM HEPES, pH 7.8, 5 mM EDTA, and 2 mM DTT) and layered in a polyallomer centrifuge tube. Mitochondrial samples were loaded on top of the gradient and ultracentrifuged at 32,000 rpm for 1 hr at 4°C. Intact organelles that banded at the interface of the sucrose cushion were carefully extracted, washed twice in SHE buffer, and stored at −80°C until LC-MS/MS or western blot analyses.

Mitochondrial Respiration

Mitochondrial respiration was determined using an XF24 Extracellular Flux Analyzer (Seahorse Bioscience) using 15 μg mitochondrial protein in a buffer containing 50 mM KCl, 4 mM KH2PO4, 5 mM HEPES, and 1 mM EGTA, 4%

BSA, 10 mM pyruvate, 5 mM malate, 1 mM GDP. Mitochondria were plated and centrifuged 2,000 g for 20 min to promote adherence to the XF24 V7 cell-culture microplate. Uncoupled and maximal OCR was determined using oligomycin (14 μM) and FCCP (10 μM). Rotenone and antimycin A (4 μM each) were used to inhibit complex 1- and complex 3-dependent respiration.

Primary Inguinal Adipocyte Differentiation

Primary inguinal preadipocytes were counted and plated in the evening at 20,000 cells per well of a seahorse plate. The following morning, inguinal preadipocytes were induced to differentiate with 1 μM rosiglitazone, 0.5 mM isobutylmethylxanthine (IBMX), 1 μM dexamethasone, and 5 μg ml^{-1} insulin. Cells were re-fed every 48 hr with 1 μM rosiglitazone and 5 μg ml^{-1} insulin. Cells were fully differentiated by day 5 post-induction.

Primary Brown Adipocyte Differentiation

Primary brown preadipocytes were counted and plated in the evening at 15,000 cells per well of a seahorse plate. The following morning, brown preadipocytes were induced to differentiate with 1 μM rosiglitazone, 0.5 mM IBMX, 5 μM dexamethasone, 0.114 μg ml^{-1} insulin, 1 nM T3, and 125 μM indomethacin. Cells were re-fed every 48 hr with 1 μM rosiglitazone and 0.5 μg ml^{-1} insulin. Cells were fully differentiated by day 5 post-induction.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.035.

AUTHOR CONTRIBUTIONS

Conceptualization, L.K. and B.M.S.; Methodology, L.K., E.T.C., B.M.S.; Investigation, L.K., E.T.C., M.P.J., B.K.E., K.S., G.Z.L.; Resources, D.L.-B., P.C., R.V., S.C.H.; Writing – Original Draft, L.K., E.T.C., B.M.S.; Writing – Review & Editing, L.K., E.T.C., P.C., B.M.S.; Funding Acquisition, L.K., E.T.C., S.K., S.P.G., B.M.S.; Supervision, S.K., S.P.G., B.M.S.

ACKNOWLEDGMENTS

We are grateful to Marc W. Kirschner, Mike P. Murphy, Bé Wieringa, and members of the Spiegelman lab for helpful discussions. pBAD-TOPO-TA/hPhospho1 was a kind gift from Colin Farquharson. The Biophysical Instrumentation Facility (NSF-0070319) is acknowledged for help with DSC. This work was supported by a Canadian Institutes of Health Research postdoctoral fellowship to L.K., by a Human Frontier Science Program postdoctoral fellowship to E.T.C., and by an NIH grant (DK031405) and the JPB foundation to B.M.S.

Received: April 26, 2015
Revised: July 10, 2015
Accepted: September 8, 2015
Published: October 22, 2015

REFERENCES

Bal, N.C., Maurya, S.K., Sopariwala, D.H., Sahoo, S.K., Gupta, S.C., Shaikh, S.A., Pant, M., Rowland, L.A., Bombardier, E., Goonasekera, S.A., et al. (2012). Sarcolipin is a newly identified regulator of muscle-based thermogenesis in mammals. Nat. Med. *18*, 1575–1579.

Berlet, H.H., Bonsmann, I., and Birringer, H. (1976). Occurrence of free creatine, phosphocreatine and creatine phosphokinase in adipose tissue. Biochim. Biophys. Acta *437*, 166–174.

Block, B.A., O'Brien, J., and Meissner, G. (1994). Characterization of the sarcoplasmic reticulum proteins in the thermogenic muscles of fish. J. Cell Biol. *127*, 1275–1287.

Bloom, J.D., Dutia, M.D., Johnson, B.D., Wissner, A., Burns, M.G., Largis, E.E., Dolan, J.A., and Claus, T.H. (1992). Disodium (R,R)-5-[2-[[2-(3-chlorophenyl)-2-hydroxyethyl]-amino] propyl]-1,3-benzodioxole-2,2-dicarboxylate (CL 316,243). A potent beta-adrenergic agonist virtually specific for beta 3 receptors. A promising antidiabetic and antiobesity agent. J. Med. Chem. *35*, 3081–3084.

Cannon, B., and Nedergaard, J. (2004). Brown adipose tissue: function and physiological significance. Physiol. Rev. *84*, 277–359.

Cohen, P., Levy, J.D., Zhang, Y., Frontini, A., Kolodin, D.P., Svensson, K.J., Lo, J.C., Zeng, X., Ye, L., Khandekar, M.J., et al. (2014). Ablation of PRDM16 and beige adipose causes metabolic dysfunction and a subcutaneous to visceral fat switch. Cell *156*, 304–316.

Feldmann, H.M., Golozoubova, V., Cannon, B., and Nedergaard, J. (2009). UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. Cell Metab. *9*, 203–209.

Fitch, C.D., and Chevli, R. (1980). Inhibition of creatine and phosphocreatine accumulation in skeletal muscle and heart. Metabolism *29*, 686–690.

Flachs, P., Rühl, R., Hensler, M., Janovska, P., Zouhar, P., Kus, V., Macek Jilkova, Z., Papp, E., Kuda, O., Svobodova, M., et al. (2011). Synergistic induction of lipid catabolism and anti-inflammatory lipids in white fat of dietary obese mice in response to calorie restriction and n-3 fatty acids. Diabetologia *54*, 2626–2638.

Gohil, V.M., Zhu, L., Baker, C.D., Cracan, V., Yaseen, A., Jain, M., Clish, C.B., Brookes, P.S., Bakovic, M., and Mootha, V.K. (2013). Meclizine inhibits mitochondrial respiration through direct targeting of cytosolic phosphoethanolamine metabolism. J. Biol. Chem. *288*, 35387–35395.

Golozoubova, V., Hohtola, E., Matthias, A., Jacobsson, A., Cannon, B., and Nedergaard, J. (2001). Only UCP1 can mediate adaptive nonshivering thermogenesis in the cold. FASEB J. *15*, 2048–2050.

Granneman, J.G., Burnazi, M., Zhu, Z., and Schwamb, L.A. (2003). White adipose tissue contributes to UCP1-independent thermogenesis. Am. J. Physiol. Endocrinol. Metab. *285*, E1230–E1236.

Grimpo, K., Völker, M.N., Heppe, E.N., Braun, S., Heverhagen, J.T., and Heldmaier, G. (2014). Brown adipose tissue dynamics in wild-type and UCP1-knockout mice: in vivo insights with magnetic resonance. J. Lipid Res. *55*, 398–409.

Jacobus, W.E., and Lehninger, A.L. (1973). Creatine kinase of rat heart mitochondria. Coupling of creatine phosphorylation to electron transport. J. Biol. Chem. *248*, 4803–4810.

Lidell, M.E., Betz, M.J., Dahlqvist Leinhard, O., Heglind, M., Elander, L., Slawik, M., Mussack, T., Nilsson, D., Romu, T., Nuutila, P., et al. (2013). Evidence for two types of brown adipose tissue in humans. Nat. Med. *19*, 631–634.

Liu, X., Rossmeisl, M., McClaine, J., Riachi, M., Harper, M.E., and Kozak, L.P. (2003). Paradoxical resistance to diet-induced obesity in UCP1-deficient mice. J. Clin. Invest. *111*, 399–407.

Lowell, B.B., S-Susulic, V., Hamann, A., Lawitts, J.A., Himms-Hagen, J., Boyer, B.B., Kozak, L.P., and Flier, J.S. (1993). Development of obesity in transgenic mice after genetic ablation of brown adipose tissue. Nature *366*, 740–742.

Meyer, C.W., Willershäuser, M., Jastroch, M., Rourke, B.C., Fromme, T., Oelkrug, R., Heldmaier, G., and Klingenspor, M. (2010). Adaptive thermogenesis and thermal conductance in wild-type and UCP1-KO mice. Am. J. Physiol. Regul. Integr. Comp. Physiol. *299*, R1396–R1406.

Oudman, I., Clark, J.F., and Brewster, L.M. (2013). The effect of the creatine analogue beta-guanidinopropionic acid on energy metabolism: a systematic review. PLoS ONE *8*, e52879.

Ricquier, D., Gaillard, J.L., and Turc, J.M. (1979). Microcalorimetry of isolated mitochondria from brown adipose tissue. Effect of guanosine-di-phosphate. FEBS Lett. *99*, 203–206.

Roberts, S.J., Stewart, A.J., Sadler, P.J., and Farquharson, C. (2004). Human PHOSPHO1 exhibits high specific phosphoethanolamine and phosphocholine phosphatase activities. Biochem. J. *382*, 59–65.

Seale, P., Bjork, B., Yang, W., Kajimura, S., Chin, S., Kuang, S., Scimè, A., Devarakonda, S., Conroe, H.M., Erdjument-Bromage, H., et al. (2008). PRDM16 controls a brown fat/skeletal muscle switch. Nature *454*, 961–967.

Shabalina, I.G., Petrovic, N., de Jong, J.M., Kalinovich, A.V., Cannon, B., and Nedergaard, J. (2013). UCP1 in brite/beige adipose tissue mitochondria is functionally thermogenic. Cell Rep. *5*, 1196–1203.

Shinoda, K., Luijten, I.H., Hasegawa, Y., Hong, H., Sonne, S.B., Kim, M., Xue, R., Chondronikola, M., Cypess, A.M., Tseng, Y.H., et al. (2015). Genetic and functional characterization of clonally derived adult human brown adipocytes. Nat. Med. *21*, 389–394.

Sipilä, I. (1980). Inhibition of arginine-glycine amidinotransferase by ornithine. A possible mechanism for the muscular and chorioretinal atrophies in gyrate atrophy of the choroid and retina with hyperornithinemia. Biochim. Biophys. Acta *613*, 79–84.

Svensson, P.A., Jernås, M., Sjöholm, K., Hoffmann, J.M., Nilsson, B.E., Hansson, M., and Carlsson, L.M. (2011). Gene expression in human brown adipose tissue. Int. J. Mol. Med. *27*, 227–232.

Terblanche, S.E., Masondo, T.C., and Nel, W. (1998). Effects of cold acclimation on the activity levels of creatine kinase, lactate dehydrogenase and lactate dehydrogenase isoenzymes in various tissues of the rat. Cell Biol. Internat. *22*, 701–707.

Ukropec, J., Anunciado, R.P., Ravussin, Y., Hulver, M.W., and Kozak, L.P. (2006). UCP1-independent thermogenesis in white adipose tissue of cold-acclimated Ucp1-/- mice. J. Biol. Chem. *281*, 31894–31908.

Véniant, M.M., Sivits, G., Helmering, J., Komorowski, R., Lee, J., Fan, W., Moyer, C., and Lloyd, D.J. (2015). Pharmacologic effects of FGF21 are independent of the "browning" of white adipose tissue. Cell Metab. *21*, 731–738.

Wakatsuki, T., Hirata, F., Ohno, H., Yamamoto, M., Sato, Y., and Ohira, Y. (1996). Thermogenic responses to high-energy phosphate contents and/or hindlimb suspension in rats. Jpn. J. Physiol. *46*, 171–175.

Watanabe, M., Yamamoto, T., Kakuhata, R., Okada, N., Kajimoto, K., Yamazaki, N., Kataoka, M., Baba, Y., Tamaki, T., and Shinohara, Y. (2008). Synchronized changes in transcript levels of genes activating cold exposure-induced thermogenesis in brown adipose tissue of experimental animals. Biochim. Biophysic. Acta *1777*, 104–112.

Watt, I.N., Montgomery, M.G., Runswick, M.J., Leslie, A.G., and Walker, J.E. (2010). Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. Proc. Natl. Acad. Sci. USA *107*, 16823–16827.

Wu, J., Boström, P., Sparks, L.M., Ye, L., Choi, J.H., Giang, A.H., Khandekar, M., Virtanen, K.A., Nuutila, P., Schaart, G., et al. (2012). Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. Cell *150*, 366–376.

Wyss, M., and Kaddurah-Daouk, R. (2000). Creatine and creatinine metabolism. Physiol. Rev. *80*, 1107–1213.

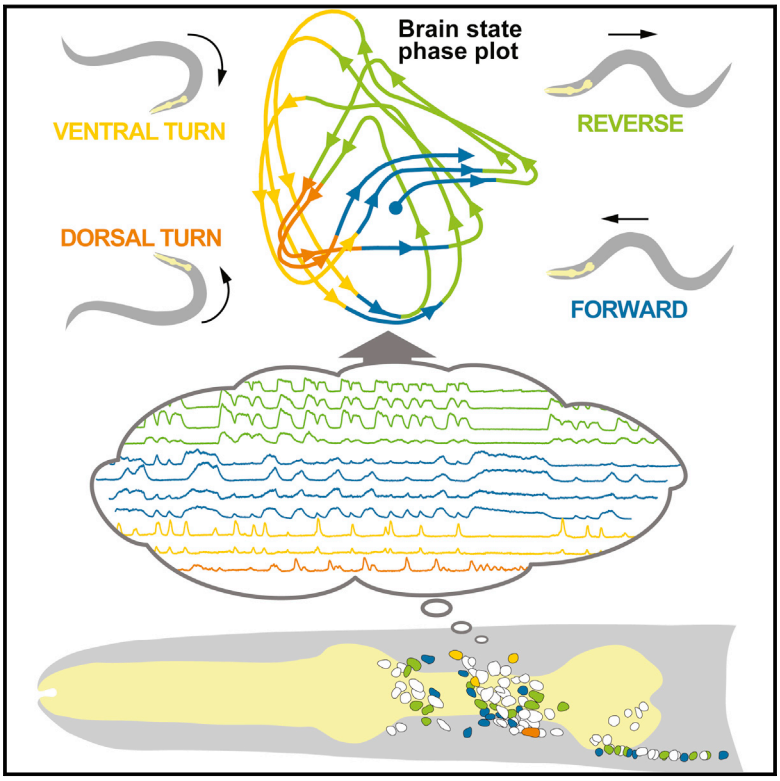
Yamashita, H., Ohira, Y., Wakatsuki, T., Yamamoto, M., Kizaki, T., Oh-ishi, S., and Ohno, H. (1995). Increased growth of brown adipose tissue but its reduced thermogenic activity in creatine-depleted rats fed beta-guanidinopropionic acid. Biochim. Biophys. Acta *1230*, 69–73.

Article

Cell

Global Brain Dynamics Embed the Motor Command Sequence of *Caenorhabditis elegans*

Graphical Abstract



Highlights

- Most active neurons in the brain participate in coordinated dynamical activity
- Smooth, cyclical dynamics continuously represent action sequences and decisions
- Internal representation of behavior persists when de-coupled from its execution
- Brain dynamics provide a robust scaffold for sensory-driven action selection

Authors

Saul Kato, Harris S. Kaplan, Tina Schrödel, ..., Eviatar Yemini, Shawn Lockery, Manuel Zimmer

Correspondence

zimmer@imp.ac.at

In Brief

Simultaneously recording the activity of nearly all neurons in the *C. elegans* brain reveals that most active neurons share information by engaging in coordinated, dynamical network activity that corresponds to the sequential assembly of motor commands.

Kato et al., 2015, Cell *163*, 1–14
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.034

CellPress

Please cite this article in press as: Kato et al., Global Brain Dynamics Embed the Motor Command Sequence of *Caenorhabditis elegans*, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.034

Article

Cell

Global Brain Dynamics Embed the Motor Command Sequence of *Caenorhabditis elegans*

Saul Kato,[1,4] Harris S. Kaplan,[1,4] Tina Schrödel,[1,4] Susanne Skora,[1] Theodore H. Lindsay,[2,5] Eviatar Yemini,[3] Shawn Lockery,[2] and Manuel Zimmer[1,*]

[1]Research Institute of Molecular Pathology IMP, Vienna Biocenter VBC, Dr. Bohr-Gasse 7, 1030 Vienna, Austria

[2]Institute of Neuroscience, University of Oregon, Eugene, OR 97403, USA

[3]Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, NY 10032, USA

[4]Co-first author

[5]Present address: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA

*Correspondence: zimmer@imp.ac.at

http://dx.doi.org/10.1016/j.cell.2015.09.034

SUMMARY

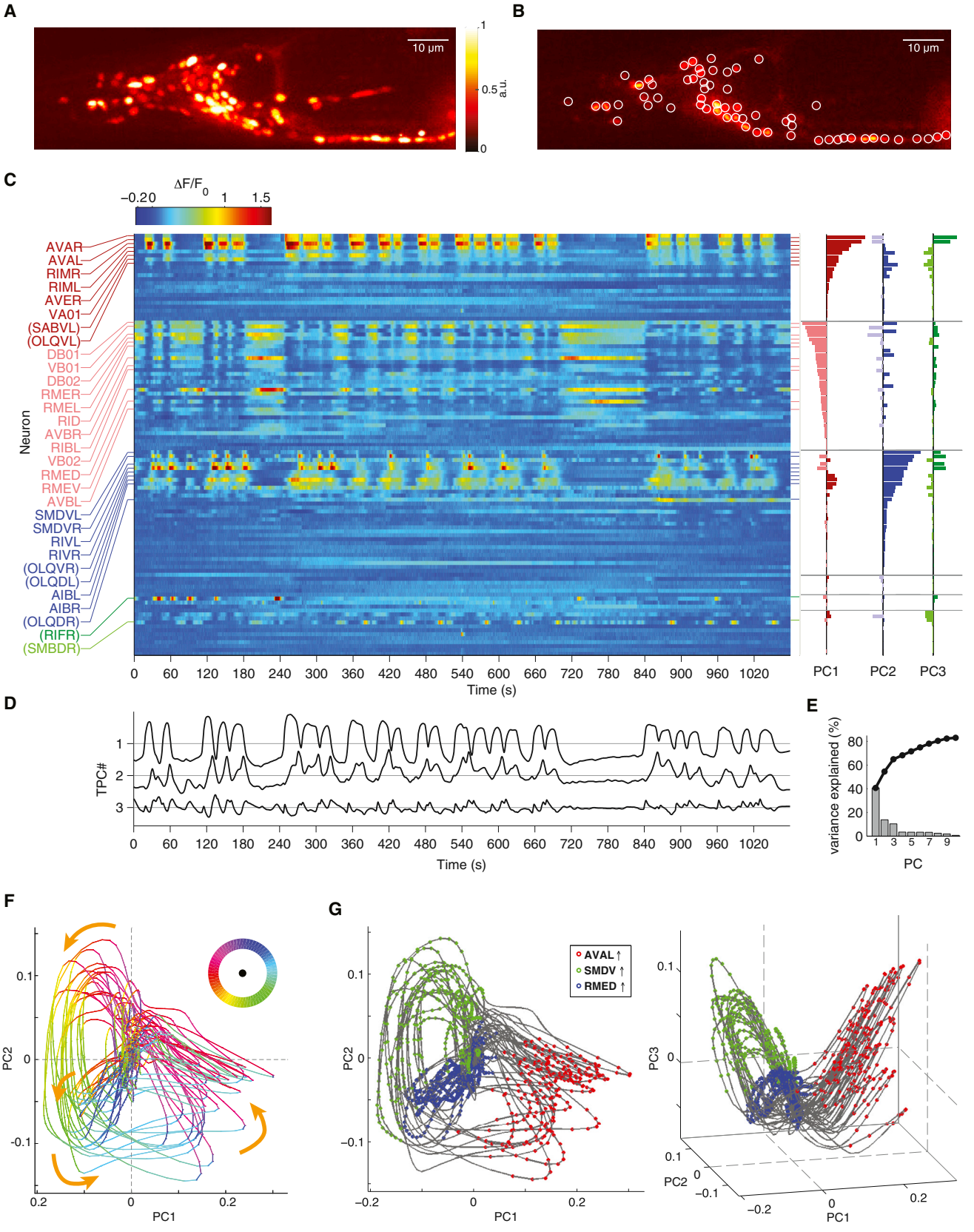
While isolated motor actions can be correlated with activities of neuronal networks, an unresolved problem is how the brain assembles these activities into organized behaviors like action sequences. Using brain-wide calcium imaging in *Caenorhabditis elegans*, we show that a large proportion of neurons across the brain share information by engaging in coordinated, dynamical network activity. This brain state evolves on a cycle, each segment of which recruits the activities of different neuronal sub-populations and can be explicitly mapped, on a single trial basis, to the animals' major motor commands. This organization defines the assembly of motor commands into a string of run-and-turn action sequence cycles, including decisions between alternative behaviors. These dynamics serve as a robust scaffold for action selection in response to sensory input. This study shows that the coordination of neuronal activity patterns into global brain dynamics underlies the high-level organization of behavior.

INTRODUCTION

Behavior is composed of individual motor actions and motifs, such as limb movements or gaits, which do not achieve organismal goals unless they are orchestrated into longer-lasting action sequences and behavioral strategies, like navigation, grooming, or courtship (Anderson and Perona, 2014; Gray et al., 2005; Seeds et al., 2014). Ethologists often make quantitative descriptions of this higher-level organization using state transition diagrams, consisting of distinct, repeatable high-level motor states and switches between them (Anderson and Perona, 2014). The brain's representation of behavior must account for both detailed metrics of individual actions (e.g., strength and extent of movement or speed of gait), as well as for their higher level orchestration. Identifying how these aspects of behavior correspond to measurable neural activity is a necessary step toward understanding how the brain encodes and produces behavior. Recent studies in invertebrate motor ganglia and mammalian cortex show that selection, execution, and shaping of motor programs correspond to neural activity patterns across large neuronal populations. These studies show that, despite the participation of hundreds of sampled neurons, their activity is coordinated, and meaningful signals can thus be reduced to far fewer dimensions. Moreover, neuronal populations encode information dynamically (Briggman et al., 2005; Bruno et al., 2015; Churchland et al., 2012; Cunningham and Yu, 2014; Harvey et al., 2012; Jin et al., 2014; Mante et al., 2013). For practical reasons, recordings in these studies have been performed over short intervals that encompass individual motions or brief behavioral tasks. Hence, the neuronal mechanisms that govern the continuous control of behavior and its time course, encompassing long-lasting and repeated action sequences, remain enigmatic. Furthermore, approaches have been typically limited by the need to average across trials or to sub-sample from local brain regions or motor ganglia. Recently, the first brain-wide single-cell-resolution functional imaging studies, in zebrafish and fly larvae and adult *C. elegans*, revealed motor-related population dynamics correlated across distant brain regions. These data suggest that behaviorally relevant neural representations might occur at the level of global population dynamics and highlight the benefit of brain-wide sampling (Ahrens et al., 2012, 2013; Lemon et al., 2015; Panier et al., 2013; Prevedel et al., 2014; Schrödel et al., 2013).

The nematode *C. elegans* is an attractive model system to address these problems, due to its stereotypic nervous system of just 302 identifiable neurons grouped into 118 anatomical symmetry classes (White et al., 1986). However, prior to the availability of whole-brain imaging, past studies had not explored distributed or population dynamics in *C. elegans*. Instead, identified interneurons and pre-motor neurons have been described as dedicated encoders of specific sensory inputs or motor outputs and are commonly placed in a context of isolated sensory-to-motor pathways (see the following references for examples: Chalasani et al., 2007; Donnelly et al., 2013; Gray et al., 2005; Ha et al., 2010; Iino and Yoshida, 2009; Kimata et al., 2012). However, these pathways largely overlap and are embedded in a horizontally organized and recurrently connected neuronal wiring diagram (Varshney et al., 2011; White et al., 1986). Moreover, recent functional imaging

Please cite this article in press as: Kato et al., Global Brain Dynamics Embed the Motor Command Sequence of *Caenorhabditis elegans*, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.034



(legend on next page)

studies revealed that many of these circuit elements encode motor rather than sensory related signals (Gordus et al., 2015; Hendricks et al., 2012; Laurent et al., 2015; Li et al., 2014; Luo et al., 2014). Taken together, these considerations argue against separable feed-forward sensory pathways and instead support the hypothesis that sensorimotor processing is performed by distributed, shared networks operating on widespread motor representations.

In the present study, we provide evidence for this hypothesis by showing that many neurons in the *C. elegans* brain participate in a pervasive dynamic population state, collectively representing the major motor commands of the animal. The time evolution of the neural state is directional and cyclical, corresponding to the sequential order of the animals' repeated actions. These network dynamics interface with sensory representations as early as at the first synapse downstream of sensory neurons and provide a robust scaffold for sensory inputs to modulate behavior. Our work suggests that high-level organization of behavior is encoded in the brain by globally distributed, continuous, and low-dimensional dynamics.

RESULTS

Brain-wide Activity Evolves on a Low-Dimensional Attractor-like Manifold

We performed whole-brain single-cell-resolution Ca^{2+} imaging with a pan-neuronally expressed nuclear Ca^{2+} sensor in animals immobilized in a microfluidic device (Schrödel et al., 2013). In each animal (n = 5), we recorded the brain activity under environmentally constant conditions for 18 min at a rate of ~2.85 volumes per second. The imaging volume spanned all head ganglia, including most of the worm's sensory neurons and interneurons, as well as all head motor neurons and the most anterior ventral cord motor neurons (White et al., 1986) (Figures 1A and 1B). In each recording, we detected 107–131 neurons and were able to determine the cell class identity of most of the active neurons. Figures 1C and S1A show a typical multi-neuron time series during which a large proportion of imaged neurons exhibited discernable Ca^{2+}-activity patterns. We performed principal components analysis (PCA) on the time derivatives of the normalized Ca^{2+} traces (Figures 1C–1E). This method produces neuron weight vectors, termed principal components (PCs); here, PCs are calculated based on the covariance structure found in the normalized data (Jolliffe, 2002). For each PC, a corresponding time series (temporal PC) was calculated by taking the weighted average of the full multi-neural time series. Temporal PCs represent signals shared by neurons that cluster based on their correlations. We found a low-dimensional, widely shared, dominant signal: the first three PCs accounted for 65% of the full dataset variance (Figure 1E). We performed PCA on the time derivatives of Ca^{2+} traces because the resulting PCs produced more spatially organized state space trajectories, described below.

The time integral of temporal PC1 displayed a strong oscillatory time course with variable period, sharp transitions, and prolonged plateaus and troughs. This pattern derived from the antagonistic activity of two groups of interneurons and motor neurons (Figure 1C, right) previously implicated in controlling the switch between forward- and backward-directed crawling (Table S1 summarizes published results). Neurons previously reported to have opposing roles were observed to have opposing signs of their PC1 weights—e.g., AVA promoting backward crawling and AVB promoting forward crawling. PC2 and PC3 received high contributions from head motor neurons. Two of these neurons (SMDV and RIV) have been implicated in postural changes required for navigational re-orientation maneuvers (termed omega turns) (Gray et al., 2005). However, the neuronal weights of all three PCs indicated contributions from many neurons (Figure 1C). PC1–3 weights and their variance contributions were consistent across the five datasets (Figures S2A–S2D).

The phase plot of temporal PC1–3 showed that the neural state's time evolution was cyclical—i.e., the same states were repeatedly revisited within a trial, such that successive trajectory cycles formed spatially coherent bundles (Figure 1F and Movie S1). Consequently, the entire neural state trajectory traced out a manifold, which is defined here as the sub-volume in PCA space occupied by the neural state trajectory. When mapped onto the neural trajectory, individual neurons' activity rise and fall phases occupied class-specific sub-regions on the manifold (Figures 1G and S1B). All five recordings displayed a similarly structured manifold (Figure S2E). Thus, a large group of interneurons and motor neurons produces a cyclical, low-dimensional population state time-varying signal.

Interneurons and Head Motor Neurons Reliably Encode Motor State and Graded Motion Parameters

Next, we aimed for a functional interpretation of the neural state manifold and its properties. Each manifold sub-region was labeled specifically and consistently by different subsets of neurons, some of which have been previously implicated in the action sequence termed a pirouette (Table S1), which is central to navigation (Gray et al., 2005; Pierce-Shimomura et al., 1999). During pirouettes, worms switch transiently from

Figure 1. Brain-wide Activity Is Organized in a Low-Dimensional, Cyclical Neural State Space Trajectory

(A) Maximum intensity projection of a representative sample recorded under constant conditions.

(B) Single z plane overlaid with segmented neuronal regions.

(C) Heat plot of fluorescence ($\Delta F/F$) time series of 109 segmented head neurons, one neuron per row. Labeled neurons indicate putative cell IDs. Ambiguous neuron IDs are in parentheses (see Figure S1 for additional candidates). Neurons are colored and grouped by their principal component (PC1–3) weights and signs, which are shown by the bar plots on the right.

(D) Integrals of the first three temporal PCs.

(E) Variance explained by first ten PCs, black line indicates cumulative variance explained.

(F) Phase plot of first two temporal PCs colored by direction of time evolution indicated by color key.

(G) Phase plots of first two (left) and first three (right) temporal PCs. Colored balls indicate Ca^{2+} rises of three example neurons indicated by legend.

See also Movie S1 and Figures S1 and S2.

forward- to backward-directed crawling, termed a reversal (Figures 2A and 2B). They then resume forward crawling with a concomitant turn along the dorsal or ventral body axis; worms crawl lying on their left or right side (Figures 2C and 2D). We performed Ca^{2+} imaging experiments of representative neurons in freely moving worms while simultaneously recording their behavior with an infrared (IR) camera (Faumont et al., 2011). We selected neurons based on their PC weights and availability of specific promoters to drive GCaMP expression. As with brain-wide imaging experiments, animals were recorded 5–10 min after removal from food, a paradigm in which pirouettes contribute to a local search strategy (Gray et al., 2005). Behavioral analysis of the IR movies showed that reversal initiations were each preceded by a reduction in crawling speed (slowing bout), though 20% of slowing bouts did not lead to a reversal (Figures S3A and S3B). We thus defined slowing as an additional behavioral state and represent pirouettes together with forward crawling as action sequences composed of forward run, slowing, reversal, resume forward via dorsal turn, and resume forward via ventral turn actions, which is depicted in a state transition diagram (Figure 2E).

We first examined Ca^{2+} dynamics in neurons with high positive or negative PC1 weight. An example trace of RIM neurons is shown in Figure 2F. We found that the Ca^{2+} signals of RIM resided in stable low states during forward-directed crawling and that Ca^{2+} rises occurred exclusively during reversals (Figure 2F). The slope of these signals correlated with the speed of reverse crawling (Figure 2G). Although reversals are of variable duration (Gray et al., 2005; Pokala et al., 2014) (Figure S3B), RIM Ca^{2+} rise onsets precisely aligned with reversal start, and RIM Ca^{2+} fall onsets aligned with reversal end. This relationship was highly reliable—approximately 90% of reversals were associated with a detectable RIM Ca^{2+} rise phase (Figure 2H, top), and the remainders were very short reversals where small Ca^{2+} signals might have been occluded by noise (Figure 2F). All clearly discernible RIM Ca^{2+} rises above our signal-to-noise threshold occurred during reversals. We found such a relationship of Ca^{2+} rise and fall phases with respect to reversal events for all tested neurons with positive PC1 weight (RIM, AVA, AVE, AIB), while neurons with negative PC1 weight (RIB, AVB, RMEV) showed the inverse relationship (Figures 2H and S3C–S3H). All these neurons' activities changed as reliably as RIM at both forward-reverse and reverse-forward transitions.

Besides this common property of PC1 neurons, class-specific relationships between neuronal activity and locomotion were revealed by freely moving Ca^{2+} imaging. RIM and AVA Ca^{2+} rise slopes, and AVE Ca^{2+} signal magnitude, were graded and correlated with reverse crawling speed (Figures 2G, S3I, and S3J). Unlike RIM, AVA, and AVE, the activity of AIB did not show strong correlations with reverse crawling speed (Figures S3E and S3K); however, small AIB Ca^{2+} transients co-occurred with forward slowing bouts, even when no reversal followed (Figures S3E and S3Q). Consistent with this, AIB Ca^{2+} rise phases preceded the forward-to-reversal transition by ~1 s on average (Figure 2H). The continuous activity of AVB and RIB, unlike RMEV, showed strong correlations with forward crawling speed (Figures S3L–S3P; see also Li et al., 2014). Consistent with this, AVB and RIB Ca^{2+} fall phases preceded the forward to reverse transition by ~1 s on average (Figure 2H).

Next, we examined the activity of SMDV head motor neurons as representative neurons with strong PC2/3 weight. Resumption of forward crawling begins with a dorsal or ventral bend, which was biased (71%/29%) in the ventral direction. The head flexure during post-reversal turns is graded and increased compared to normal forward crawling, especially for ventral bends (Figure 3A). SMDV exhibited Ca^{2+} rises at the transition from reverse to forward crawling; importantly, these rises occurred exclusively during ventrally and not dorsally directed events (Figures 3B–3D). The magnitude of these signals correlated with ventral head-bending flexure (Figure 3E).

The major qualitative divergence in neural activity patterns between the freely moving single neuron and restrained whole-brain setups that we observed was the absence, in freely moving worms, of prolonged high phases in neurons with positive PC1 weight. Using RIM as an exemplar, we first ruled out that this difference was a consequence of nuclear localization of the Ca^{2+} reporter used in whole-brain imaging (Figures S3R–S3T). We then dissociated the two major differences in these experimental conditions by performing experiments in either pharmacologically or physically immobilized worms. While low doses of the paralyzing agent tetramisole caused RIM high phases in conjunction with prolonged slowly executed reversals, physical immobilization alone also caused RIM high phases (Figures S3U–S3X). These data suggest that impeded motor execution leads to a prolongation of the reversal, which is correlated with sustained Ca^{2+} levels in reversal-promoting neurons.

In summary, the investigated neuronal activities showed both (1) sharp transitions depending on discrete motor state (i.e., forward versus backward crawling, ventral versus dorsal turning direction) and (2) graded information about motion parameters (i.e., forward and reverse crawling speed and head bending flexure). Acute motor state reliably matched the activities of the associated neurons on a single event basis. Importantly, when examining neuron activity periods mapped onto the neural state manifold, we observed that neurons encoding the same behavioral state in freely moving animals shared the same manifold sub-regions with rare exception (Figure S1B).

Manifold Branches and Bundles Exhibit Distinct Neuronal Recruitment Patterns

Having determined that the neural state manifold is a composite of motor related signals, we next aimed for a quantitative description thereof. We first segmented the global brain cycle into four behaviorally relevant phases using the left AVA neuron (AVAL) as a reference: a trough in AVAL Ca^{2+} defined the LOW state, a Ca^{2+} increase the RISE state, a Ca^{2+} plateau the HIGH state, and a Ca^{2+} decrease the FALL state (Figure 4A). We chose this single neuron class because it is among the highest PC1 contributors, participated in every brain cycle, and, unlike temporal PCs, exhibited sharply discernible transitions; however, other strongly PC1-contributing neurons such as RIM could also be used for this purpose. We validated that the appearance of lasting plateau and smooth transition states was not due to temporal filtering effects of Ca^{2+} imaging: all four states were readily discernible in AVA membrane voltage recordings, and

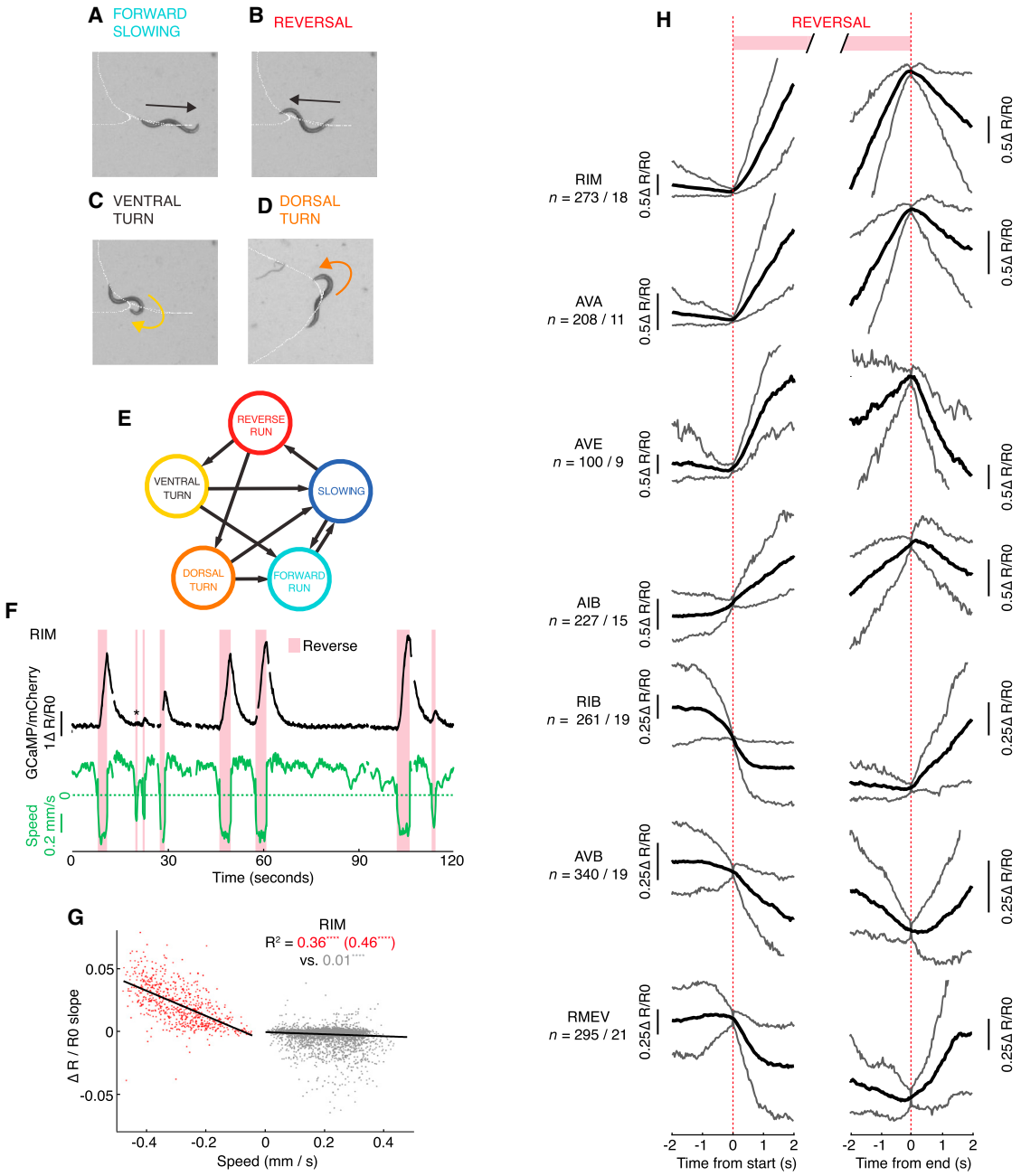


Figure 2. Distributed Encoding of Motor State and Crawling Speed by Interneurons in Freely Moving Worms

(A–D) Motor states of pirouette action sequence. White dotted lines show crawling trajectory. Arrows indicate crawling direction.

(E) Behavioral state transition diagram indicating motor states as circles and possible transitions as arrows.

(F–H) Ca^{2+} imaging in freely moving animals.

(F) Example trace showing RIM activity as normalized GCaMP/mCherry fluorescence ratio (black) and corresponding crawling speed (green). Pink bars overlay reverse crawling periods. Asterisk indicates reversal with no detectable RIM activity peak.

(G) Regression analysis of crawling speed versus RIM Ca^{2+} signal slope. R^2 indicates goodness of linear fit for instantaneous and maximum (in parentheses) reverse speed (red) and instantaneous forward speed (gray). Permutation test p value ****$p < 0.0001$ indicates probability that correlation was obtained by chance.

(H) Average Ca^{2+} signals of the indicated neurons triggered to reversal start (left) or end (right). Upper and lower traces represent 90th and 10th percentile of all data, respectively. Number of recorded worms and reversal events are indicated.

See also Figure S3.

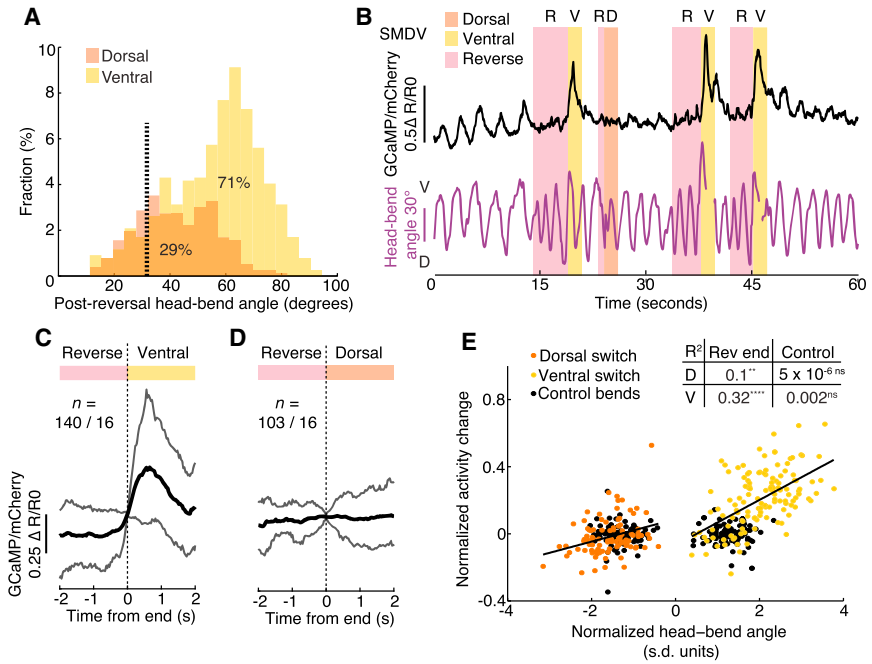


Figure 3. SMDV Signals during Ventral, but Not Dorsal, Post-Reversal Turns

(A) Fractional histogram showing postural angle of first post-reversal head bend (ventral, yellow; dorsal, orange). Numbers indicate percentage of all post reversal head bends. Dashed vertical black line shows median of all other head bends (no difference between ventral and dorsal).

(B–E) SMDV Ca^{2+} imaging in freely moving animals.

(B) Example trace showing SMDV activity as normalized GCaMP/mCherry fluorescence ratio (black) and corresponding head-bend angle (purple). Pink bars overlay reverse crawling; yellow and orange bars overlay ventral and dorsal post-reversal head-bends, respectively.

(C and D) Average SMDV Ca^{2+} signals triggered to reversals ending with ventral (C) or dorsal (D) head bends. Upper and lower traces represent 90th and 10th percentile of all data, respectively. Number of recorded worms and events are indicated.

(E) Regression analysis of normalized peak post-reversal head-bend angle versus SMDV Ca^{2+} signal. Ventral and dorsal bends are shown in yellow and orange, respectively. Black open circles show an equal number of randomly selected head-bend peaks during regular forward movement. R^2 indicates goodness of linear fits to ventral (V), dorsal (D), and respective control groups. Permutation test p values (****$p < 0.0001$, **$p < 0.01$, ^{ns} not significant) indicate probability that R^2 value was obtained by chance.

we calculated an estimate of low-pass filtering caused by nuclear Ca^{2+}-imaging, producing a maximum delay in signal peaks of less than 1.1 s (Figure S4). Although neurons with a common relationship to behavior were recruited to the same sub-regions of the manifold, their precise phase onsets and offsets varied. In order to quantify this observation, for each onset of RISE and FALL, we created a vector containing the phase delays of all recruited neurons (Figure S5) (see Supplemental Experimental Procedures for details). Across the five datasets, we detected 121 RISE and 123 FALL transitions and observed characteristic phase delay distributions for each neuronal class (Figure S5). Next, we searched for structure across neuronal classes by performing k-means clustering separately for the RISE and FALL phase timing vectors; we found that both could be significantly clustered into two groups each, which we termed RISE1/2 and FALL1/2, respectively. RISE1 differed from RISE2 mostly based on different timing of neurons; e.g., AIB and RIB activity exhibited phase advances during RISE1 (Figure S5). FALL1 and FALL2 mostly differed by mutually exclusive head motor neuron recruitments, SMDV/RIV versus RMED/ventral ganglion head motor neuron (likely SMB, SMDD, or RMF) (Figure S5). The precise ordering detected by this method may be affected by differential Ca^{2+} dynamics in different cells; however, the reproducible clustering would be preserved. Using this six-state classification (LOW, RISE1/2, HIGH, and FALL1/2), we labeled the neural state trajectory and found that each state classifies a distinct bundle of trajectory segments (Figures 4A and 4B and Movie S2). Thus, the two methods (PCA and phase timing analysis) revealed the same dynamical structure in the neural data. Bundle classification enabled us to calculate average neural state trajectories illustrating the canonical brain cycle (Figure 4C). Note that, without this single-trial clustering analysis, the cycle-averaged trajectory would be reduced to a single loop in neural state space. Furthermore, bundle classification enabled us to estimate a contour surface of the manifold (Figure 4D and Movie S3), where the extents correspond to the standard deviations (SDs) by which the trajectory path diverges from the canonical (average) path. The trajectory segments across all cycles are strongly bundled; the mean pairwise distance of points across any two phase-registered trajectory time points within a bundle is ~10% of the diameter of the full trajectory, and their mean angular divergence is 22° versus 90° expected from uncorrelated orientations. In summary, we find that many active neurons across the brain are tightly bound to reproducible and smooth population dynamics.

The Motor Command Sequence Is Embedded in Neural State Space

Remarkably, the relationships neurons exhibited with behavioral transitions (Figures 2H, 3C, and 3D) matched their phase relationships with the six state global brain cycle without exception. Assembling all of the neuronal-behavioral correlate information gathered via Ca^{2+} imaging in freely moving worms enabled us to unambiguously map the worm's major motor command states onto separate bundles of the neural state manifold (Figures 4B–4E)—RISE1 or RISE2, in conjunction with HIGH, correspond to reversals, with HIGH corresponding to the sustained reversal seen only in immobilized animals. FALL1 corresponds to the post-reversal ventral turn and FALL2 to the dorsal turn. FALL1 and FALL2, in conjunction with LOW, correspond to forward crawling. Slowing mapped to final sections of LOW

preceding RISEs (Figures 4B–4E, see Experimental Procedures for the detailed mapping rules). Thus, the neural state manifold, on a single trial basis, embeds the pirouette command sequence described in the state transition diagram (Figures 2A–2E). The neural trajectory follows the same unidirectional sequence through manifold sub-regions as the corresponding behavioral sequence executed by freely moving worms during pirouettes. This observation motivated us to redraw the state transition diagram (Figure 2E) as a continuous flow graph (Figure 4E). The neuronal manifold, in addition to embedding the command sequence, also contains information about graded locomotion parameters like the drive underlying crawling speed (Figure 4F, see Experimental Procedures for the detailed mapping rules). Both motor command states, as well as speed drive, appear organized on the manifold; i.e., separable sub-regions unambiguously delimit the distinct command states (Figure 4B) and proximal traversals on the manifold exhibit similar speed drives (Figure 4F). This manifold organization was clearly apparent in all five recordings (Figure S2E).

Each branching region of the manifold represents a decision where the subsequent motor state is determined. To explore the process of decision execution, we quantified the time course of trajectory separation when branching into RISE1 versus RISE2 and FALL1 versus FALL2 and subsequent merging. This approach calculates how significantly trajectory segments bundle in PCA space when tested against random shuffling of membership in RISE1 versus RISE2 or FALL1 versus FALL2 clusters (see Supplemental Experimental Procedures for details). Consistent with the significant clustering of neuronal recruitment vectors described above, there was significant separation during the RISE and FALL phases (Figures 4G and 4H). Interestingly, this also uncovered memory effects: a RISE1 versus RISE2 branch choice could, on average, be predicted during the preceding FALL period (Figure 4G), and consistent with the previous, FALL1 versus FALL2 trajectories remained significantly unmixed in the following RISE phases (Figure 4H). Moreover, RISE1 and RISE2 are associated, respectively, with long and short preceding LOW states (Figure 4I). Both results indicate that the trajectory path history influences the future branch choice decision.

In contrast to the state transition diagram, the neural state manifold captures the continuous dynamical structure of motor commands and their transitions and contains additional information about graded metrics of motion, like crawling speed and postural flexure. Here, we define the terms command state and speed drive as the brain's internal high-level representations of the underlying motor programs, since these are readily observable in immobilized animals in the absence of motor execution.

Neural State Dynamics Persist When a Hub Output Neuron Is Inhibited

The presence of a representation of the pirouette sequence in immobilized animals suggests that the neuronal population dynamics are primarily internally driven and thus represent descending motor commands that can operate in the absence of motor feedback. We sought to further test this hypothesis. Despite the largely recurrent connectivity of the *C. elegans* wiring diagram, a bottleneck exists from the head ganglia to body motor neurons—AVA pre-motor interneurons are anatomical network hubs linking head ganglia neurons to A-class ventral cord motor neurons, which mediate the reversal motor program (Chalfie et al., 1985; Kawano et al., 2011; Varshney et al., 2011). Acutely silencing AVA via transgenic expression of a histamine-gated chloride channel (HisCl) (Pokala et al., 2014) abolished reversals in freely moving worms (Figure 5A). As expected, similarly silenced animals under whole-brain imaging (n = 5 recordings) showed substantial attenuation of AVA activity and strong uncoupling of AVA from the global brain cycle (Figures 5B and S6A). Additionally, activity of the reverse interneurons AVE and RIM, which are connected to AVA via gap junctions (White et al., 1986) was slightly attenuated (Figure 5B). However, their phase relationships with most other neurons appeared normal (Figure S6C). A-class ventral cord motor neurons, the principal output targets of AVA, also showed significant attenuation (Figure 5B). Despite these effects, the cyclical dynamics and neuronal recruitment patterns were largely preserved (Figures 5C, 5D, and S6). The distributions of network state durations were unchanged, with the exception of a decrease in HIGH state duration, suggesting that network HIGH state prolongation was due in part to reinforcement from AVA (Figure 5E). These observations raised the possibility that the global brain cycle was also intact in freely moving worms with AVA, and therefore reversals, inhibited. Unlike in wild-type animals, where 92.5% of turns occurred in conjunction with a preceding reversal, in worms with silenced AVA neurons, none of the turns were preceded by reversals; instead, 68% of turns (32 out of 47) were preceded by prolonged slowing or pauses, while the rest occurred during apparently normal forward locomotion. Imaging RIM in AVA-silenced freely moving animals revealed the presence of sustained RIM activity during these prolonged slowing or pauses preceding normal turning events (Figures 5F–5H). Such transients were never seen in controls, where RIM was only active during reversals. In AVA-silenced animals, RIM activity often entered HIGH states during prolonged pauses, further supporting the above interpretation that the HIGH state occurs due to the absence of effectual motor execution (Figures 5F and S3U–S3X). These results show that the cyclical time course of the brain-wide motor command is maintained in the absence of reversal execution, the only effect of which is a prolonged HIGH state duration. Analogously, behaviors that are not AVA-output mediated (slowing and turns) are also preserved. Further, these data imply that AVA is not a privileged generator of motor commands but should instead be characterized as an output-facing member of the collectively oscillating interneuron group.

Entrainment of the Global Brain Cycle by Sensory Stimulation

Next, we investigated how these collective network dynamics interact with a chemosensory input. Under whole-brain imaging, we stimulated oxygen chemosensory neurons with consecutive oxygen up- and down-shifts (21% versus 4%), a protocol previously shown to reliably activate BAG, URX, and AQR oxygen sensory neurons and to entrain pirouette behavior with high pirouette probability at 21% oxygen and low at 4% (Figures S7A and S7B; see also references Busch et al., 2012; Schrödel et al., 2013; Zimmer et al., 2009). To our surprise, with the exception of one ventral ganglion neuron class (RIG or RIF)

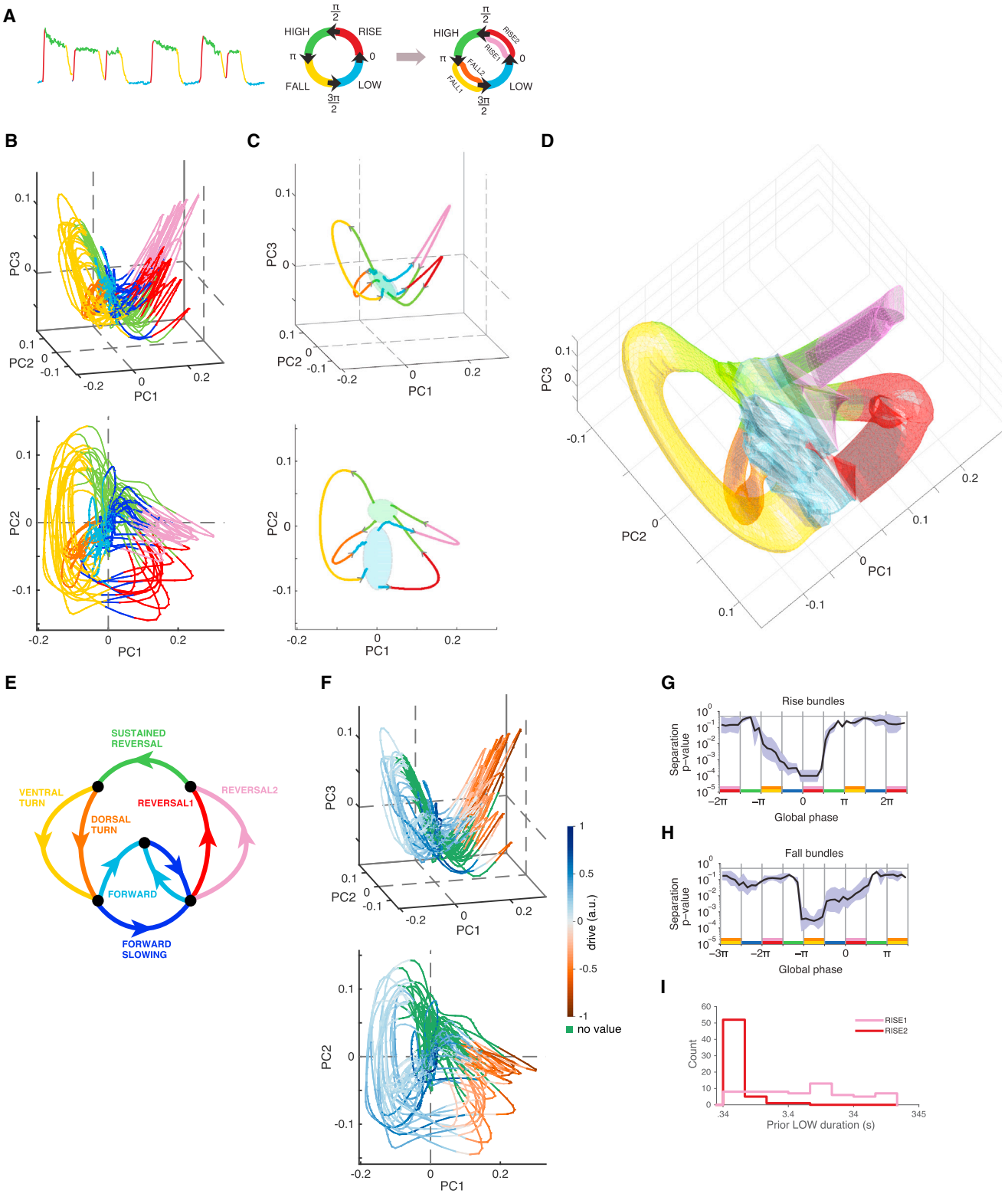


Figure 4. The Neural State Manifold Embeds the Action Sequence and Exhibits Organized Analog Speed Drive

(A) Phase segmentation of example AVAL trace (left). Four-state brain cycle (middle). Phase timing analysis and clustering leads to six-state brain cycle (right). See also Figures S4 and S5.

(B) Phase plot of the same trial shown in Figure 1, colored by six-state brain cycle plus FORWARD SLOWING command state in purple (see below).

(legend continued on next page)

(Figure S7C), we did not detect single-neuron representations of sensory stimulus downstream of sensory neurons (n = 13 recordings). Moreover, the topology of the neural state manifold did not change upon stimulation; however, there were some magnitude effects on the amplitude of temporal PC1 (Figure 6A). Based on the strong entrainment effect the stimulation protocol has on pirouette behavior, we expected that oxygen concentration should affect bundle occupancy on the manifold. Indeed, the stimulus protocol entrained the global phase of the brain cycle so that the probability of the reverse motor command state declined during 4% oxygen periods and increased during 21% oxygen periods (Figures 6B and 6C), indicating a successful sensorimotor transformation in our preparation. Consistent with these findings, Ca^{2+} rises in BAG neurons during the HIGH state evoked immediate FALL1 or FALL2 transitions in 56% (30/54, n = 13 recordings) of all instances (see Figure S7C for an example). Interestingly, in 22 out of the 24 remaining instances, secondary BAG Ca^{2+}-rises coincided with a FALL1 or FALL2 transition; these were the only times when we observed secondary BAG transients (see Figure S7C as an example). This finding suggests the existence of a feedback mechanism eliciting or gating secondary Ca^{2+} rises in the BAG sensory neurons, demonstrating that variability in the BAG sensory response profile (Zimmer et al., 2009) can be explained when the underlying brain state is known to the observer.

Finally, we looked for sensory-evoked Ca^{2+} activity in the major PC1 neuron classes AVA, AVE, and RIB in freely moving animals. Together AVE and RIB receive 47% of BAG neuron synapses (White et al., 1986). Consistent with our whole-brain imaging results, these neurons retained a tight correlation with motor state and movement metrics and lacked obvious sensory encoding activity; the magnitude of Ca^{2+} signals was subtly modulated during the stimulation periods (Figures S7D–S7U).

In summary, neural state manifold organization is robust to a salient sensory input and thus stably encodes the motor command sequences of the worm under these conditions. The major effect of sensory input was to modulate the probability that the neural state resides on a particular segment bundle by driving the neural state along a lawful trajectory. The result is an entrainment of the global brain cycle, which is consistent with the entrainment of corresponding motor behaviors in freely moving worms.

DISCUSSION

In this work, we identify and characterize a brain-wide signal in *C. elegans* that dominates the neural activity time series. Although our approach required the use of a nuclear localized Ca^{2+} indicator, omitting the detection of subcellular Ca^{2+} signals (Chalasani et al., 2007; Hendricks et al., 2012; Li et al., 2014), it reveals a pervasive motor state representation that is shared among most interneuron and motor neuron layers. The neural state trajectory exhibits directional, cyclical flow (Figure 1F) confined to a low-dimensional manifold (Figure 4D), organized into bundles (Figures 4B–4D) composed of stereotyped and smoothly changing neural activity vectors (Figure S5). Each motor command within the pirouette action sequence is reliably represented across several neurons. Neurons additionally encode graded parameters of locomotion, e.g., crawling speed and postural flexure (Figures 2, 3, and S3). These data enable us to unambiguously map behavioral commands onto sub-regions of the neural state manifold, enabling instantaneous behavioral decoding throughout an experimental trial (Figures 4B and 4E). We interpret these dynamics as corresponding to motor commands, as they can be decoupled from motor output either by restraint (during whole-brain imaging) or manipulation of a major output neuron (Figure 5). Organized flow along the neural state manifold mediates the assembly of motor commands into action sequences (Figures 4B and 4E); it thus represents the high-level temporal organization of behavior upstream of the generation of the animal's undulatory gait. This contrasts with population dynamics in the motor ganglia of crustaceans, mollusks, and lampreys that generate peristaltic and movement rhythms (Bruno et al., 2015; Grillner, 2006; Marder and Bucher, 2007). Interestingly, the brain's forward and reversal motor commands are coupled to corresponding rise, high, fall, and low states in the B- and A-class ventral nerve cord (VNC) motor neurons (Figures 1 and S1), which is consistent with previous studies performed in moving *C. elegans*. Additionally, VNC motor neuron activity exhibits gait-related rhythmic activity superimposed on these command states (Kawano et al., 2011; Wen et al., 2012), which requires proprioceptive coupling to movement (Wen et al., 2012). Taken together, we propose that behavioral state is encoded in the brain and coupled to the motor periphery and that this coupling co-occurs with locally maintained rhythmic activity.

These continuous neural dynamics embed behavioral motifs, described by the state transition diagram, and permit their superposition with graded motion metrics (Figure 4F). The process of decision making leading to execution of alternate behaviors can be observed as the time evolution of neural trajectories before the branches (Figures 4B–4D, 4G, and 4H). We propose that the phenomenon of global dynamics robustly and continuously encoding action sequence commands may be present in

(C) Phase-registered averages of the two RISE phase and two FALL phase bundles colored by six-state brain cycle. Semi-transparent ovals denote trajectory bundle mixing regions.
(D) Contour surface illustrating the neural state manifold colored by six-state brain cycle.
(E) Flow diagram indicating the motor command states corresponding to the six-state brain cycle plus FORWARD SLOWING command state (purple).
(F) The same phase plot colored by forward- and reverse-speed drive inferred from neural correlate decoding. Green trajectory segments indicate the SUSTAINED REVERSAL state, for which no drive correspondence is made. See Figure S2 for more examples.
(G and H) Quantification of inter-bundle separation and mixing for RISE (G) and FALL (H) clusters. Traces show trial-averaged p values (shading indicates SEM; n = 5 animals) of mean normalized pairwise distance at instantaneous points in the past or future, which indicate the probability that the observed separation between bundles occurred by chance. This calculation was done in six dimensions (PC1–3 plus their derivatives) to incorporate directional information from the trajectory paths.
(I) Distribution of LOW state durations preceding RISE1 or RISE2 segments.
See also Movies S2 and S3.

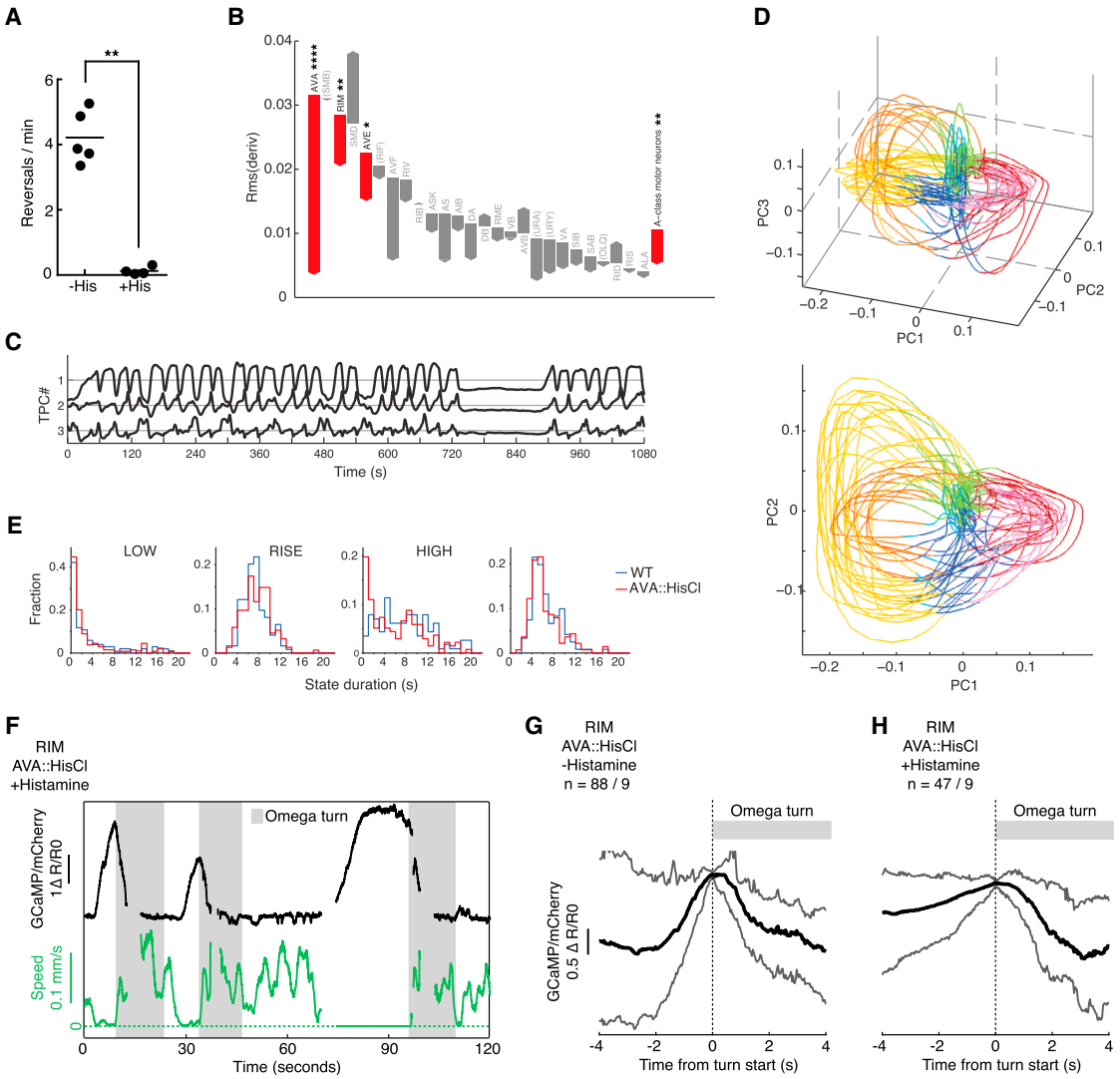


Figure 5. Global Brain Dynamics Persist when Decoupled from Motor Output

(A) Reversal events per minute for AVA::HisCl worms without (−His) or with (+His) histamine treatment. Each data point represents a single assay, n = 20–25 worms per assay. Horizontal lines show means. Mann-Whitney test, $^{**}p < 0.01$.

(B) Shifts in trial-averaged root-mean-squared power of neuronal trace derivatives of AVA::HisCl worms with histamine treatment, relative to wild-type control (n = 5). Gray bars indicate non-significant power shifts, red bars indicate significant power shifts. Class-A motor neurons, typically 1–2 visible per recording, were combined. Significance was determined using a permutation test, $^{****}p < 0.0001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

(C and D) Integrated temporal PCs (C) and phase plots (D) of an example AVA::HisCl dataset.

(E) Distributions of state durations of AVA::HisCl (red) versus wild-type (blue) across multiple trials (n = 5).

(F–H) Ca^{2+} imaging of RIM in freely moving animals expressing HisCl in AVA.

(F) Example trace showing RIM activity in an AVA::HisCl worm after histamine treatment. Normalized GCaMP/mCherry fluorescence ratio (black) and corresponding crawling speed (green) are shown. Omega turns are indicated with gray overlaid bars. These worms did not exhibit reversals.

(G and H) Averages of RIM Ca^{2+} signals in AVA::HisCl worms triggered to omega turn onset, for worms pre-incubated without (G) or with (H) histamine. Upper and lower traces represent 90th and 10th percentile of all data, respectively. Number of recorded worms and omega turns are indicated.

See also Figure S6.

higher animals with more sophisticated behavioral repertoires. This hypothesis is supported by the observation of smooth population dynamics maintaining navigational plans in rodents (Harvey et al., 2012). Its generality could be further tested by studying the basis of well-described sequential courtship and grooming behaviors in fruit flies (Dankert et al., 2009; Seeds et al., 2014). The ability to find dynamical structure solely on the basis of neural event timing (Figure S5) suggests that the structure we observe is not a particular consequence of the graded, non-spiking, nature of *C. elegans* neurons. We speculate that neuronal population trajectories associated with action selection in leeches (Briggman et al., 2005), limb movement in monkey

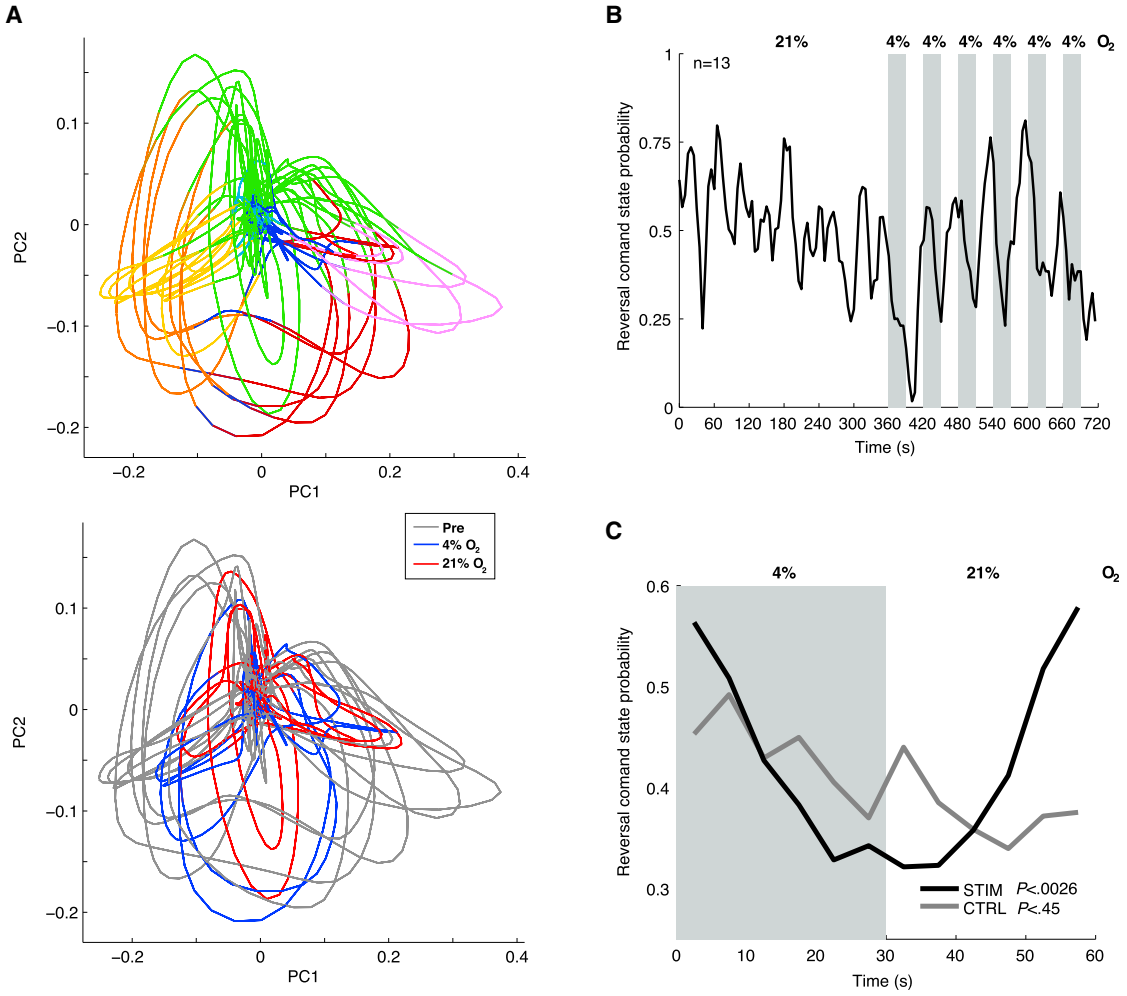


Figure 6. Entrainment of the Global Brain Cycle by Sensory Stimulation

Animals were recorded and stimulated with the oxygen profile indicated in (B).

(A) Phase plots of temporal PCs 1–2 from a representative recording. Top: behavioral command state coloring as in Figure 4B. Bottom: trajectory segments during the pre-stimulus period are labeled gray; segments during the 4% and 21% shift periods are labeled blue and red, respectively.

(B) The trace shows the probability of reversal command state (REVERSAL1 + REVERSAL2 + SUSTAINED REVERSAL) calculated over n = 13 recordings.

(C) Reversal command state probability as in (B) but averaged over the six down- and up-shift periods. p values are calculated by a resampling test and indicate the probability that the stimulus-synced profile shape occurred from a randomly time-shifted stimulus pattern.

See also Figure S7.

cortical areas (Georgopoulos and Carpenter, 2015; Shenoy et al., 2013), and speech in humans (Bouchard et al., 2013) may be sparsely sampled windows onto similarly well-organized, smooth global dynamics.

Our work establishes a framework for future studies aimed at embedding more fine-scaled behaviors beyond the discrete classifications of the state transition diagram, such as gradual steering commands (Iino and Yoshida, 2009) and locomotory gait (Stephens et al., 2008). By exploring more sophisticated sensory input paradigms and studying the animal in different contexts and life stages, we expect that the neural state manifold will be further sub-dividable and support the mapping of other behavioral parameters. Additionally, in-depth analysis of whole-brain activity may uncover previously hidden aspects of behavior; for example, we found two types of reversals (corresponding to RISE1 and RISE2) in whole-brain activity that currently lack known behavioral correlates. Although AVA inhibition had only subtle effects, systematically expanding this approach to other neurons and combinations thereof should reveal whether individual neurons or sub-ensembles are causal to brain dynamics. By probing the system with acute perturbation using optogenetics and imaging at finer timescales and sub-neuronal spatial resolution, it should be possible to uncover the neuronal logic governing trajectory control and branch selection, which underlies decision making in this system. Measuring manifold geometry changes over longer timescales may uncover the characteristics of brain states such as hunger-satiety or sleep-wakefulness.

Our results argue against models of largely feed-forward sensory-to-motor flow where intermediate neuronal layers perform sequential processing and the behavioral state is only ultimately represented within the nervous system at the motor periphery. Instead, our data support a model of an early interface between sensory and motor representations as was suggested by recent single-neuron studies (Hendricks et al., 2012; Luo et al., 2014).

Moreover, motor command representations affect responsiveness of sensory neurons and early interneurons to sensory inputs via feedback mechanisms (Figure S7) that remain to be identified (see also Gordus et al., 2015). Consistent with recent distributed models of sensorimotor action selection in mammals, including primates (Cisek and Kalaska, 2010), our work suggests that the brain's outputs—i.e., its intents and actions—make up a large fraction of its dynamic activity state.

Our findings reveal that a large collection of neuronal classes with distinct morphologies and connectivities (White et al., 1986), distinct molecular compositions and neurotransmitter expression patterns (Hobert, 2013), distinct synaptic transmission properties (Li et al., 2014), and distinct subcellular signal processing capacities (Chalasani et al., 2007; Hendricks et al., 2012; Li et al., 2014) nevertheless collectively share a low-dimensional, pervasive neuronal signal. The class-specific phase relationships with respect to the global brain cycle (Figures S1B and S5) suggest that neurons differentially interact with this shared mode. We therefore propose that the neural state manifold influences and binds local activity to a global reference framework, establishing a consensus that produces stable, coherent behavior.

EXPERIMENTAL PROCEDURES

The Supplemental Experimental Procedures contain more detailed information on each procedure, and in addition, they include descriptions of region of interest detection and neural time series extraction from volumetric Ca^{2+} imaging data, electrophysiology, simulation of nuclear GCaMP signals from voltage traces, population behavior assays, statistics applied in this study, strain genotypes, and molecular biology constructs.

Whole-Brain Ca^{2+} Imaging of *C. elegans* Head Ganglia Neurons

Animals were immobilized with 1 mM tetramisole in microfluidic devices that allow controlled O_2 stimuli as previously described (Schrödel et al., 2013; Zimmer et al., 2009). Recordings were started within 5 min after removal from food. Worms were either imaged for 18 min at constant 21% O_2 or, for the stimulus protocol, imaged for 12 min with the first 6 min at 21% O_2 and the remaining 6 min with 30 s consecutive shifts between 4% and 21% O_2. Data were acquired using an inverted spinning disc microscope (UltraViewVoX, PerkinElmer) equipped with an EMCCD camera (C9100-13, Hamamatsu).

Identification of Head Ganglia Neurons

In each recording, we detected 107–131 neurons, covering 55%–67% of expected neurons in the imaging area. Neurons were identified taking into account their anatomical positions, also in relation to surrounding neurons (http://www.wormatlas.org), and their activity patterns. To confirm ambiguous neuron identities, marker lines expressing red fluorophores in neurons of interest were generated and crossed to the imaging line expressing GCaMP5K pan-neuronally in the nucleus (ZIM504).

Time Series Analysis: PCA, Numerical Differentiation, 4-Phase Segmentation, Phase Timing Analysis, and Clustering

PCA was performed on the time derivatives of $\Delta F/F_0$ neural traces, each normalized by its peak magnitude. To compute de-noised time derivatives without the need of smoothing that can affect precise timing of sharp transitions, the total-variation regularization method (Chartrand, 2011) was applied. To segment individual neuronal activity into 4-phase sequences, first RISE and FALL phases for neurons were identified as periods when the time derivative was greater or lower than a small threshold, respectively. HIGH and LOW phases were then inferred in the remaining gaps. For trajectory segment averaging (Figures 4C, 4D, and S2E) and generation of Movies S2 and S3, neuronal time series were registered to a common phase clock by matching phase segment starts and ends to the reference neuron (AVA or RIM) rise onsets and fall offsets, respectively, followed by linearly interpolating within phase segments. To perform phase timing analysis, first a set of global transitions, either RISE or FALL onsets, were defined by the transitions of a reference neuron (AVA or RIM in this study). Then, relative time delays of the nearest transitions found in other neurons were used to compose a feature vector for each global transition. In the absence of a matching transition within 7 s of the reference neuron transition, a time delay of −10 s was used for the purposes of clustering, since the absence of neurons was also considered an important feature of transitions. *K*-means clustering was applied to transition feature vectors for each full trial using L_1 distance and $k = 2$. Detailed explanations of the above computational analyses may be found in the Supplemental Experimental Procedures.

Behavioral Decoding of Whole-Brain Recordings

Each time point of the phase plot trajectory was first assigned to a global brain cycle HIGH, LOW, RISE1, RISE2, FALL1, or FALL2 segment as described above and in the main text, then mapped to motor command states as follows. RISE1 and RISE 2 segments were mapped to REVERSAL1 and REVERSAL2 command states, respectively. HIGH segments were mapped to the SUSTAINED REVERSAL state. FALL1 and FALL2 segments were mapped to VENTRAL TURN and DORSAL TURN, respectively. LOW segments were mapped to FORWARD except that RIB FALL phases present during global LOW segments were mapped to FORWARD SLOWING command states. A speed drive was assigned to each point on the trajectory as follows, aside from those in SUSTAINED REVERSAL phases for which no speed drive was inferred. During VENTRAL TURN, DORSAL TURN, FORWARD, and FORWARD SLOWING phases, positive speed drive was taken to be the magnitude of RIB activity, normalized to its most negative value during the trial. During REVERSAL1 and 2 phases, negative speed drive was taken to be the derivative of RIM neuron activity, normalized to its highest value during the trial.

Ca^{2+} Imaging in Freely Moving Animals

Ca^{2+} imaging recordings were made using the automatic re-centering system described previously (Faumont et al., 2011) with custom modifications. Young adult worms (0–8 eggs) expressed both mCherry and GCaMP in the neuron of interest. Animals were recorded while freely crawling on agar in a custom built microscope stage containing an airtight chamber with inlet and outlet connectors for gas flow delivery. Images were acquired using two CCD cameras (Evolve 512, Photometrics) connected via a DualCam DC2 beam splitter (Photometrics). A long-distance 63× objective (Zeiss LD Plan-Neofluar 63×, 0.75 NA) was used to obtain unbinned images streamed at 30.3 frames per second (fps) acquisition rate. Simultaneous behavior recordings under infrared illumination (780 nm) were made using a CCD camera (Manta Prosilica GigE, Applied Vision Technologies) at 4× magnification and 10 fps acquisition rate.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, one table, and three movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.034.

AUTHOR CONTRIBUTIONS

S.K. designed experiments, developed analytical methods for whole-brain imaging datasets, and analyzed data. H.S.K. designed experiments, generated transgenic strains, performed Ca^{2+}-imaging experiments in freely moving animals, developed analytical methods, and analyzed data. T.S. designed

experiments, generated transgenic strains, performed whole-brain imaging experiments, and analyzed data. S.S. performed population behavioral recordings and analyzed data. T.H.L. and S.L. performed electrical recordings; E.Y. wrote code for behavioral analysis; and M.Z. designed experiments, developed analytical methods, and led the project. S.K., H.S.K., T.S., and M.Z. wrote the manuscript.

ACKNOWLEDGMENTS

We thank Cori Bargmann, Larry Abbott, Alipasha Vaziri, Andrew Straw, Hagai Lalazar, Omri Barak, and Sean Escola for critically reading the manuscript, Richard Latham for technical support, and Martin Colombini for manufacturing of mechanical components. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 281869 (acronym: *elegans Neurocircuits*) to M.Z., the Simons Foundation (grant number 324958 to M.Z.), an EMBO Long Term Fellowship to S.K. (number ALTF 345-2014), an NIH training grant to T.H.L. (number F31 NS061697), an NIH T32 training grant to E.Y. (number T32 MH015174-38), and the Research Institute of Molecular Pathology (IMP). The IMP is funded by Boehringer Ingelheim.

Received: July 2, 2015
Revised: August 14, 2015
Accepted: September 2, 2015
Published: October 15, 2015

REFERENCES

Ahrens, M.B., Li, J.M., Orger, M.B., Robson, D.N., Schier, A.F., Engert, F., and Portugues, R. (2012). Brain-wide neuronal dynamics during motor adaptation in zebrafish. Nature *485*, 471–477.

Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., and Keller, P.J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods *10*, 413–420.

Anderson, D.J., and Perona, P. (2014). Toward a science of computational ethology. Neuron *84*, 18–31.

Bouchard, K.E., Mesgarani, N., Johnson, K., and Chang, E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature *495*, 327–332.

Briggman, K.L., Abarbanel, H.D., and Kristan, W.B., Jr. (2005). Optical imaging of neuronal populations during decision-making. Science *307*, 896–901.

Bruno, A.M., Frost, W.N., and Humphries, M.D. (2015). Modular deconstruction reveals the dynamical and physical building blocks of a locomotion motor program. Neuron *86*, 304–318.

Busch, K.E., Laurent, P., Soltesz, Z., Murphy, R.J., Faivre, O., Hedwig, B., Thomas, M., Smith, H.L., and de Bono, M. (2012). Tonic signaling from O_2 sensors sets neural circuit activity and behavioral state. Nat. Neurosci. *15*, 581–591.

Chalasani, S.H., Chronis, N., Tsunozaki, M., Gray, J.M., Ramot, D., Goodman, M.B., and Bargmann, C.I. (2007). Dissecting a circuit for olfactory behaviour in Caenorhabditis elegans. Nature *450*, 63–70.

Chalfie, M., Sulston, J.E., White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1985). The neural circuit for touch sensitivity in Caenorhabditis elegans. J. Neurosci. *5*, 956–964.

Chartrand, R. (2011). Numerical differentiation of noisy, nonsmooth data. ISRN Applied Mathematics *2011*, 1–11.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Cisek, P., and Kalaska, J.F. (2010). Neural mechanisms for interacting with a world full of action choices. Annu. Rev. Neurosci. *33*, 269–298.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500–1509.

Dankert, H., Wang, L., Hoopfer, E.D., Anderson, D.J., and Perona, P. (2009). Automated monitoring and analysis of social behavior in Drosophila. Nat. Methods *6*, 297–303.

Donnelly, J.L., Clark, C.M., Leifer, A.M., Pirri, J.K., Haburcak, M., Francis, M.M., Samuel, A.D.T., and Alkema, M.J. (2013). Monoaminergic orchestration of motor programs in a complex C. elegans behavior. PLoS Biol. *11*, e1001529.

Faumont, S., Rondeau, G., Thiele, T.R., Lawton, K.J., McCormick, K.E., Sottile, M., Griesbeck, O., Heckscher, E.S., Roberts, W.M., Doe, C.Q., and Lockery, S.R. (2011). An image-free opto-mechanical system for creating virtual environments and imaging neuronal activity in freely moving Caenorhabditis elegans. PLoS ONE *6*, e24666.

Georgopoulos, A.P., and Carpenter, A.F. (2015). Coding of movements in the motor cortex. Curr. Opin. Neurobiol. *33*, 34–39.

Gordus, A., Pokala, N., Levy, S., Flavell, S.W., and Bargmann, C.I. (2015). Feedback from network states generates variability in a probabilistic olfactory circuit. Cell *161*, 215–227.

Gray, J.M., Hill, J.J., and Bargmann, C.I. (2005). A circuit for navigation in Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA *102*, 3184–3191.

Grillner, S. (2006). Biological pattern generation: the cellular and computational logic of networks in motion. Neuron *52*, 751–766.

Ha, H.I., Hendricks, M., Shen, Y., Gabel, C.V., Fang-Yen, C., Qin, Y., Colón-Ramos, D., Shen, K., Samuel, A.D.T., and Zhang, Y. (2010). Functional organization of a neural network for aversive olfactory learning in Caenorhabditis elegans. Neuron *68*, 1173–1186.

Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

Hendricks, M., Ha, H., Maffey, N., and Zhang, Y. (2012). Compartmentalized calcium dynamics in a C. elegans interneuron encode head movement. Nature *487*, 99–103.

Hobert, O. (2013). The neuronal genome of Caenorhabditis elegans (WormBook), pp. 1–106.

Iino, Y., and Yoshida, K. (2009). Parallel use of two behavioral mechanisms for chemotaxis in Caenorhabditis elegans. J. Neurosci. *29*, 5370–5380.

Jin, X., Tecuapetla, F., and Costa, R.M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. Nat. Neurosci. *17*, 423–430.

Jolliffe, I.T. (2002). Principal Component Analysis, Second Edition (Springer).

Kawano, T., Po, M.D., Gao, S., Leung, G., Ryu, W.S., and Zhen, M. (2011). An imbalancing act: gap junctions reduce the backward motor circuit activity to bias C. elegans for forward locomotion. Neuron *72*, 572–586.

Kimata, T., Sasakura, H., Ohnishi, N., Nishio, N., and Mori, I. (2012). Thermotaxis of C. elegans as a model for temperature perception, neural information processing and neural plasticity. Worm *1*, 31–41.

Laurent, P., Soltesz, Z., Nelson, G.M., Chen, C., Arellano-Carbajal, F., Levy, E., and de Bono, M. (2015). Decoding a neural circuit controlling global animal state in C. elegans. eLife *4*, 4.

Lemon, W.C., Pulver, S.R., Höckendorf, B., McDole, K., Branson, K., Freeman, J., and Keller, P.J. (2015). Whole-central nervous system functional imaging in larval Drosophila. Nat. Commun. *6*, 7924.

Li, Z., Liu, J., Zheng, M., and Xu, X.Z.S. (2014). Encoding of both analog- and digital-like behavioral outputs by one C. elegans interneuron. Cell *159*, 751–765.

Luo, L., Wen, Q., Ren, J., Hendricks, M., Gershow, M., Qin, Y., Greenwood, J., Soucy, E.R., Klein, M., Smith-Parker, H.K., et al. (2014). Dynamic encoding of perception, memory, and movement in a C. elegans chemotaxis circuit. Neuron *82*, 1115–1128.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 1–19.

Marder, E., and Bucher, D. (2007). Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. Annu. Rev. Physiol. *69*, 291–316.

Panier, T., Romano, S.A., Olive, R., Pietri, T., Sumbre, G., Candelier, R., and Debrégeas, G. (2013). Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. Front. Neural Circuits *7*, 65.

Pierce-Shimomura, J.T., Morse, T.M., and Lockery, S.R. (1999). The fundamental role of pirouettes in Caenorhabditis elegans chemotaxis. J. Neurosci. *19*, 9557–9569.

Pokala, N., Liu, Q., Gordus, A., and Bargmann, C.I. (2014). Inducible and titratable silencing of Caenorhabditis elegans neurons in vivo with histamine-gated chloride channels. Proc. Natl. Acad. Sci. USA *111*, 2770–2775.

Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E.S., and Vaziri, A. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. Nat. Methods *11*, 727–730.

Schrödel, T., Prevedel, R., Aumayr, K., Zimmer, M., and Vaziri, A. (2013). Brain-wide 3D imaging of neuronal activity in Caenorhabditis elegans with sculpted light. Nat. Methods *10*, 1013–1020.

Seeds, A.M., Ravbar, P., Chung, P., Hampel, S., Midgley, F.M., Jr., Mensh, B.D., and Simpson, J.H. (2014). A suppression hierarchy among competing motor programs drives sequential grooming in Drosophila. eLife *3*, e02951.

Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. Annu. Rev. Neurosci. *36*, 337–359.

Stephens, G.J., Johnson-Kerner, B., Bialek, W., and Ryu, W.S. (2008). Dimensionality and dynamics in the behavior of C. elegans. PLoS Comput. Biol. *4*, e1000028.

Varshney, L.R., Chen, B.L., Paniagua, E., Hall, D.H., and Chklovskii, D.B. (2011). Structural properties of the Caenorhabditis elegans neuronal network. PLoS Comput. Biol. *7*, e1001066.

Wen, Q., Po, M.D., Hulme, E., Chen, S., Liu, X., Kwok, S.W., Gershow, M., Leifer, A.M., Butler, V., Fang-Yen, C., et al. (2012). Proprioceptive coupling within motor neurons drives C. elegans forward locomotion. Neuron *76*, 750–761.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci. *314*, 1–340.

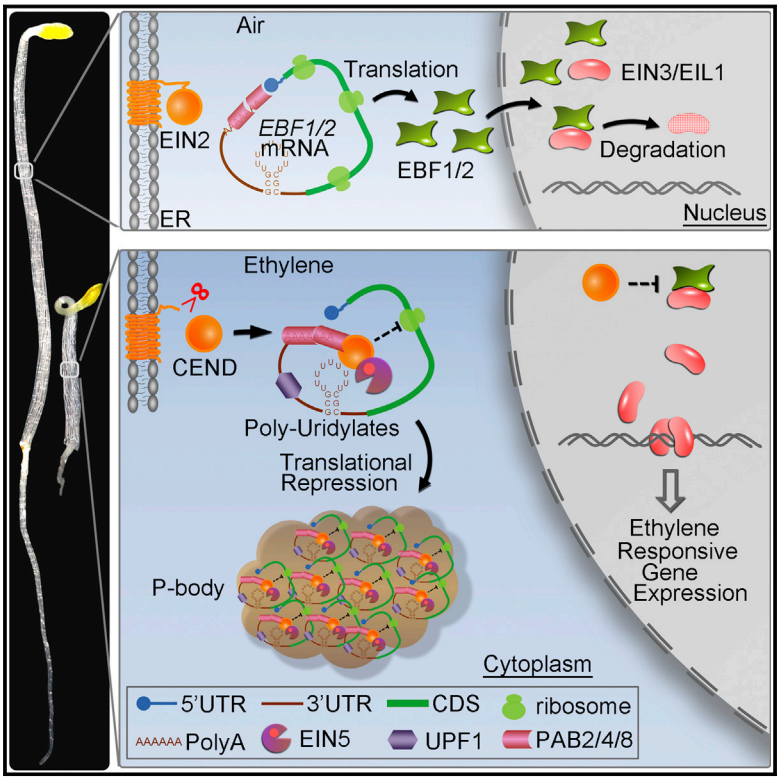
Zimmer, M., Gray, J.M., Pokala, N., Chang, A.J., Karow, D.S., Marletta, M.A., Hudson, M.L., Morton, D.B., Chronis, N., and Bargmann, C.I. (2009). Neurons detect increases and decreases in oxygen levels using distinct guanylate cyclases. Neuron *61*, 865–879.

Article

Cell

EIN2-Directed Translational Regulation of Ethylene Signaling in *Arabidopsis*

Graphical Abstract



Highlights

- Ectopic expression of *EBF1/2* 3′ UTR fragment leads to ethylene insensitivity
- 3′ UTR mediates ethylene-induced translational repression in an EIN2-dependent way
- PolyU motifs within 3′ UTR are critical for EIN2-directed translational inhibition
- EIN2 targets *EBF1* 3′ UTR to cytoplasmic P-body via interacting with EIN5 and PABs

Authors

Wenyang Li, Mengdi Ma, Ying Feng, ..., Mingzhe Li, Fengying An, Hongwei Guo

Correspondence

hongweig@pku.edu.cn

In Brief

This study reports a novel translational repression mechanism during ethylene signaling in which 3′ UTRs of mRNAs function as signal transducers.

Li et al., 2015, Cell *163*, 670–683
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.037

CellPress

Cell

Article

EIN2-Directed Translational Regulation of Ethylene Signaling in *Arabidopsis*

Wenyang Li,[1,3] Mengdi Ma,[1,3] Ying Feng,[1] Hongjiang Li,[1,4] Yichuan Wang,[1] Yutong Ma,[1] Mingzhe Li,[1] Fengying An,[1,2] and Hongwei Guo[1,2,*]

[1]The State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China

[2]Peking-Tsinghua Center for Life Sciences, Beijing 100871, China

[3]Co-first author

[4]Present address: Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria

*Correspondence: hongweig@pku.edu.cn

http://dx.doi.org/10.1016/j.cell.2015.09.037

SUMMARY

Ethylene is a gaseous phytohormone that plays vital roles in plant growth and development. Previous studies uncovered EIN2 as an essential signal transducer linking ethylene perception on ER to transcriptional regulation in the nucleus through a "cleave and shuttle" model. In this study, we report another mechanism of EIN2-mediated ethylene signaling, whereby EIN2 imposes the translational repression of *EBF1* and *EBF2* mRNA. We find that the *EBF1/2* 3′ UTRs mediate EIN2-directed translational repression and identify multiple poly-uridylates (PolyU) motifs as functional *cis* elements of 3′ UTRs. Furthermore, we demonstrate that ethylene induces EIN2 to associate with 3′ UTRs and target *EBF1/2* mRNA to cytoplasmic processing-body (P-body) through interacting with multiple P-body factors, including EIN5 and PABs. Our study illustrates translational regulation as a key step in ethylene signaling and presents mRNA 3′ UTR functioning as a "signal transducer" to sense and relay cellular signaling in plants.

INTRODUCTION

Ethylene is a gaseous phytohormone produced by plants in response to various internal and environmental stimuli, which triggers a wide range of physiological and morphological responses (Johnson and Ecker, 1998). During the past decades, a relatively linear ethylene signaling pathway has been established through the application of molecular and genetic approaches (Guo and Ecker, 2004). In *Arabidopsis*, ethylene is perceived by a group of ER-located receptors (Chang and Stadler, 2001). In the absence of ethylene signal, the hormone-free receptors activate a Raf-like protein kinase CONSTITUTIVE TRIPLE RESPONSE 1 (CTR1) (Gao et al., 2003; Kieber et al., 1993). Activated CTR1 and the receptors cooperatively inhibit an ER-located membrane protein ETHYLENE INSENSITIVE 2 (EIN2) through physical interaction and protein phosphorylation (Alonso et al., 1999; Bisson and Groth, 2011; Ju et al., 2012).

EIN2 is a key component in ethylene signaling pathway, evidenced by completely ethylene-insensitive phenotypes of the *ein2* null mutants (Ji and Guo, 2013). It is encoded by a single-copy gene in *Arabidopsis*, and is conserved from charophyte green algae to land plants (Ju et al., 2015). While the function of its N-terminal membrane-spanning domain is not clear, the C-terminal end of EIN2 (CEND) is thought to participate in signaling output, as ectopic expression of this domain alone can partially activate ethylene responses (Alonso et al., 1999; Wen et al., 2012). Recent studies reported that CEND can be phosphorylated by the receptors-activated CTR1 in the absence of ethylene (Ju et al., 2012; Qiao et al., 2012). Upon ethylene application, inactivation of the receptors and CTR1 abolishes the phosphorylation state of CEND, leading to its proteolysis from the ER-tethered N terminus, followed by shuttling into the nucleus (Ju et al., 2012; Qiao et al., 2012; Wen et al., 2012). However, this "cleave and shuttle" mode might represent part of the EIN2 actions, as induced nuclear localization of CEND only partially activates ethylene signaling (Ji and Guo, 2013; Wen et al., 2012). Meanwhile, ethylene also induces CEND to form discrete and prominent foci in the cytoplasm (Qiao et al., 2012; Wen et al., 2012), but the function of such cytoplasmic portion remains unexplored.

In the nucleus, components working downstream of EIN2 are two master transcription factors ETHYLENE INSENSITIVE 3 (EIN3) and its homolog EIN3-LIKE 1 (EIL1), which regulate the vast majority of ethylene-directed gene expression (Chang et al., 2013; Chao et al., 1997). One of the key regulatory mechanisms of ethylene signaling is the stabilization of EIN3/EIL1 proteins, wherein ethylene acts to repress the proteasomal degradation of EIN3/EIL1 mediated by two F-box proteins, EIN3-BINDING F-BOX 1 (EBF1) and EBF2, in an EIN2-dependent manner (An et al., 2010; Guo and Ecker, 2003; Potuschak et al., 2003). However, the molecular mechanism of how ethylene or EIN2 represses the function of *EBF1/2* is still elusive.

ETHYLENE INSENSITIVE 5 (*EIN5*), encoding a cytoplasmic 5′-3′ exoribonuclease (AtXRN4), is another component positively modulating ethylene responses (Olmedo et al., 2006; Potuschak et al., 2006). Currently, little is known about how EIN5 modulates ethylene signaling, except for the genetic evidence suggesting its participation in the regulation of *EBF1/2* function (Olmedo et al., 2006; Potuschak et al., 2006). Notably, small RNA fragments corresponding to *EBF1* and *EBF2* mRNA 3′ UTR were processed and accumulated in *ein5* (Olmedo et al., 2006; Potuschak et al., 2006; Souret et al., 2004). Our recent work uncovered that

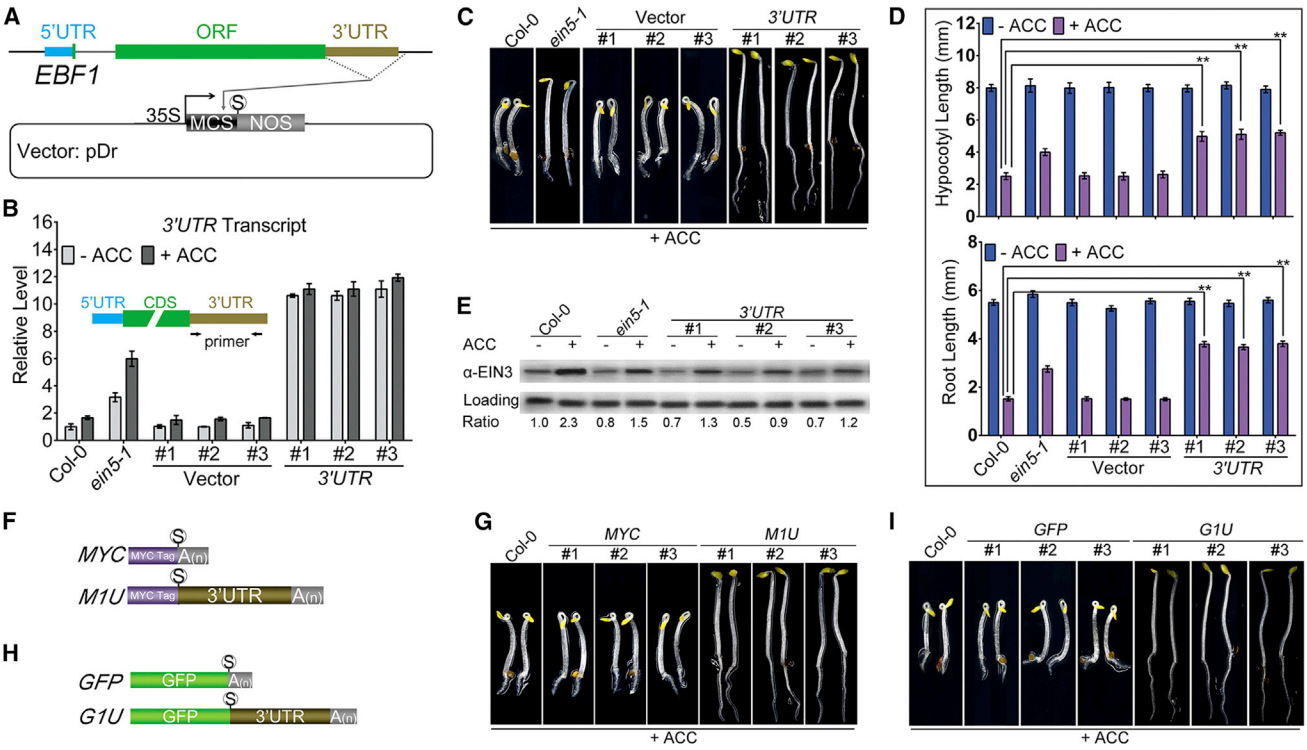


Figure 1. Overexpression of *EBF1 3′ UTR* Results in Reduced Ethylene Sensitivity

(A) Schematic diagrams of the gene structure of *EBF1* and the *3′-UTR*-overexpressing construct. Full-length *EBF1 3′ UTR* (643 bp after stop codon) plus a 66-bp flanking sequence was inserted into the multiple cloning site (MCS) prior to the NOS terminator in pDr vector. S in open circle, stop codon.

(B) Quantification of *3′ UTR* transcripts in etiolated seedlings of three independent transgenic lines grown on MS medium with (+) or without (–) ACC (an ethylene biosynthetic precursor). Vector means pDr-expressing transgenic plants while *3′ UTR* means *3′-UTR*-overexpressing transgenic lines. Arrows denote the primers used for qRT-PCR to detect the levels of *3′ UTR*.

(C) The triple response phenotypes of seedlings corresponding to (B).

(D) Quantification of hypocotyl lengths and root lengths of the seedlings in (C). **$p < 0.01$. Mean ± SD, $n > 10$.

(E) Immunoblot assays showing EIN3 protein levels of seedlings corresponding to (B). A nonspecific band served as a loading control. The numbers indicate the relative EIN3 protein levels as calculated from three biological replicates.

(F and H) Schematic maps of *M1U* (*MYC-EBF1 3′ UTR*) and *G1U* (*GFP-EBF1 3′ UTR*), as well as two control transcripts *MYC* and *GFP*. A(n) represents the poly(A) tail. Of note, all these transcripts are driven by CaMV 35S promoters.

(G and I) The triple response phenotypes of etiolated seedlings of wide-type Col-0 as well as three independent lines of indicated transgenic plants.

See also Figure S1.

EIN5, in combination with 3′-5′ RNA decay pathway, is responsible for the removal of many defective coding transcripts as well as the cleavage fragments of miRNA targets, including 3′ UTRs, which are otherwise subjected to posttranscriptional gene silencing (Zhang et al., 2015). However, genetic evidence disfavored the possibility that 3′ UTR fragments of *EBF1/2* mRNA are processed and targeted to small RNA-mediated gene silencing pathway (Potuschak et al., 2006). Interestingly, ectopic expression of a 3′ UTR-truncated *EBF2* gene resulted in a stronger ethylene insensitive phenotype than that of the *EBF2* full-length gene (Konishi and Yanagisawa, 2008), implying a negative role of 3′ UTR on the *EBF2* function.

In this study, we sought to investigate the regulatory mechanisms of how ethylene signal is relayed from cytoplasm to nucleus, and how EIN2 and EIN5 participate in this signaling process. Strikingly, we found that ectopic expression of either *EBF1* or *EBF2* 3′ UTR fragments confers strong ethylene-insensitivity phenotypes through promoting the translation of endogenous *EBF1/2* mRNAs. Furthermore, we found that ethylene induces EIN2 to target *EBF1* 3′ UTR to cytoplasmic processing-body (P-body) through interacting with EIN5 and other P-body factors to repress *EBF1/2* translation. Our study uncovers another branch of ethylene signaling pathway mediated by cytoplasmic EIN2 in translational control.

RESULTS

Overexpression of *EBF1 3′ UTR* Leads to Reduced Ethylene Sensitivity

Previous studies revealed that the *ein5* mutant accumulated *EBF1/2* mRNA 3′ UTR fragments (Olmedo et al., 2006; Potuschak et al., 2006; Souret et al., 2004). We thus speculated that the over-accumulated 3′ UTR fragments could contribute to the ethylene insensitivity of *ein5*. To test this speculation, we overexpressed the *EBF1* 3′ UTR region (*1U*) in wild-type Col-0 plants (Figures 1A and 1B). The so-called "triple response"

phenotype is commonly used as an ethylene-specific growth response in *Arabidopsis*, which refers to exaggerated apical hooks, shortened hypocotyls and roots of dark-grown seedlings exposed to ethylene or treated with ethylene precursor 1-aminocyclopropane-1-carboxylic acid (ACC) (Bleecker et al., 1988; Ecker, 1995). Overexpression of *1U* conferred significant attenuation of triple response phenotypes to Col-0, resulting in elongated hypocotyls and roots compared with control seedlings (Figures 1C and 1D). Consistently, we found that the levels of EIN3 protein were lower in *1U* transgenic plants than that in Col-0 (Figure 1E).

Furthermore, we fused *1U* to the *MYC* tag and *GFP* coding sequence (referred to as *M1U* and *G1U*), respectively (Figures 1F and 1H), and overexpressed these fusion genes in wild-type Col-0 (Figures S1A–S1C and S1F–S1H). Similar to *1U*-overexpressing seedlings, *M1U*- and *G1U*-overexpressing plants displayed reduced ethylene sensitivity and impaired EIN3 protein accumulation compared with control plants (Figures 1G, 1I, S1D, S1E, S1I, and S1J). Together, these results demonstrate that overexpression of *1U*, alone or in fusion with unrelated transcripts, reduces ethylene sensitivity.

Overexpression of *EBF1 3′ UTR* Promotes the Translation of Endogenous *EBF1/2* mRNAs

Interestingly, we found that ethylene hyposensitivity resulting from *1U*-overexpression was partially restored by a defect in either *EBF1* or *EBF2* (Figure 2A). Due to the fatal effect of over-accumulated EIN3 in *ebf1 ebf2* double mutant, we next overexpressed *M1U* in *β-estradiol-inducible EIN3-Flag/ein3 eil1 ebf1 ebf2* (*iEIN3/qm*) (An et al., 2010), which was used as a substitution of the lethal *ebf1 ebf2* double mutant (Figure S2A). We found that *M1U* no longer affected the triple response phenotypes (Figure 2B), and the abundance of EIN3 protein was comparable between *iEIN3/qm* and *M1U iEIN3/qm* (Figure S2B). Together, these results demonstrate that the presence of *EBF1/2* is required for the *1U*-overexpression-induced repression of ethylene responses, implying that exogenous *3′ UTR* expression modulates the function of *EBF1/2*.

We found that the levels of both *EBF1* and *EBF2* transcripts were not evidently affected by *1U* overexpression (Figure S2C), excluding the modulation of *EBF1/2* at the level of transcription or RNA decay. We next examined whether the translation of *EBF1/2* mRNAs is under the regulation. Without good antibody against EBF1 or EBF2 available, two experiments were conducted for this purpose. Using polysome profiling assays, we found that the translation of *EBF1* and *EBF2* mRNAs was repressed by ethylene, as the portion of high-density polysome-associated *EBF1/2* mRNAs was decreased upon ethylene application (Figure 2C). Notably, *1U* overexpression recovered the drop of the portion of polysome-associated *EBF1/2* mRNAs (Figure 2C). Therefore, *1U* overexpression augments the translation of endogenous *EBF1/2* mRNAs, which is subjected to repression by ethylene.

Furthermore, we constructed transgenic plants harboring *GFP-EBF1* followed by *1U* or not (*G1F* and *G1C*, respectively) (Figure 2D), and expressed an inducible *1U* (*iEBF1U*) in these plants to examine the effect of exogenous *1U* expression on *G1F* or *G1C* translation. We found that, while *GFP-EBF1* mRNA levels were comparable, GFP-EBF1 protein levels were downregulated by ethylene and upregulated by *1U* overexpression in *G1F* plants (Figures 2E and 2F). By contrast, in *G1C* plants, the GFP-EBF1 protein levels were virtually unchanged upon *1U* expression regardless of ethylene application (Figures 2E and 2F). Collectively, these results suggest that the over-accumulation of *1U* transcripts boosts the function of *EBF1/2* by enhancing their translation.

Based on these observations, we propose a translational interference model, in which ectopically expressed *1U* transcripts interferes with the endogenous *EBF1/2 3′ UTRs* that supposedly exert a repressive role on the translation of *EBF1/2* mRNAs. Such translational interference could arise from the competition and/or titration of translational repressors binding to the endogenous 3′ UTR regions (Figure 2G).

The 3′ UTRs Impart Translational Inhibition to *EBF1/2* mRNAs in Response to Ethylene

We next tested the translational interference model (Figure 2G) by examining the effect of *EBF1* 3′ UTR on *GFP* mRNA translation (Figures 1H, 1I, and S1F–S1J). We found that, with the comparable transcript levels (Figure S1G), seedlings expressing *G1U* accumulated much lower GFP fluorescence or protein abundance than those expressing *GFP* alone, particularly when treated with ACC (Figures 3A–3D). Ethylene caused over 80% of decrease in the translational efficiency of *G1U* whereas had no effect on *GFP* alone (Figures 3C and 3D). The ACC-promoted reduction in GFP protein abundance was restored by the application of ethylene inhibitor silver ions (Ag^+) (Figure 3E). Taken together, these results indicate that *EBF1* 3′ UTR confers translational repression to its fusion mRNA in response to ethylene.

Next, we determined the biological significance of the *EBF1* mRNA 3′ UTR–mediated translational repression in ethylene signal transduction. We constitutively expressed *M1C* (*MYC-EBF1, MYC* tag fused with the *EBF1* coding sequence) and *M1F* (*MYC* fused with the *EBF1* full-length transcript including coding sequence and 3′ UTR) (Figure 3F). Compared with control plants, *M1F* expression resulted in reduced ethylene sensitivity, whereas *M1C* expression conferred nearly complete ethylene insensitivity (Figure 3G). In agreement with the triple response phenotype, the amount of MYC-EBF1 protein was nearly constant in *M1C* but progressively decreased in *M1F* upon treatment with increasing doses of ACC (Figure 3H). Given the comparable mRNA abundance between *M1F* and *M1C* (Figures S3A and S3B), we concluded that translational repression of *EBF1* mRNA via its 3′ UTR is critical for *EBF1* function in ethylene signaling.

We further found that the overexpression of *EBF2 3′ UTR* (*2U*) also led to reduced ethylene sensitivity in *GFP-EBF2 3′ UTR* (*G2U*) transgenic plants (Figures S3C and S3D). Like *EBF1* 3′ UTR, *EBF2* 3′ UTR also conferred translational repression to the *GFP* mRNA fused with it (Figure S3E). Thus, the 3′ UTRs of both *EBF1* and *EBF2* act similarly to impose translational repression to their respective mRNAs in response to the ethylene signal.

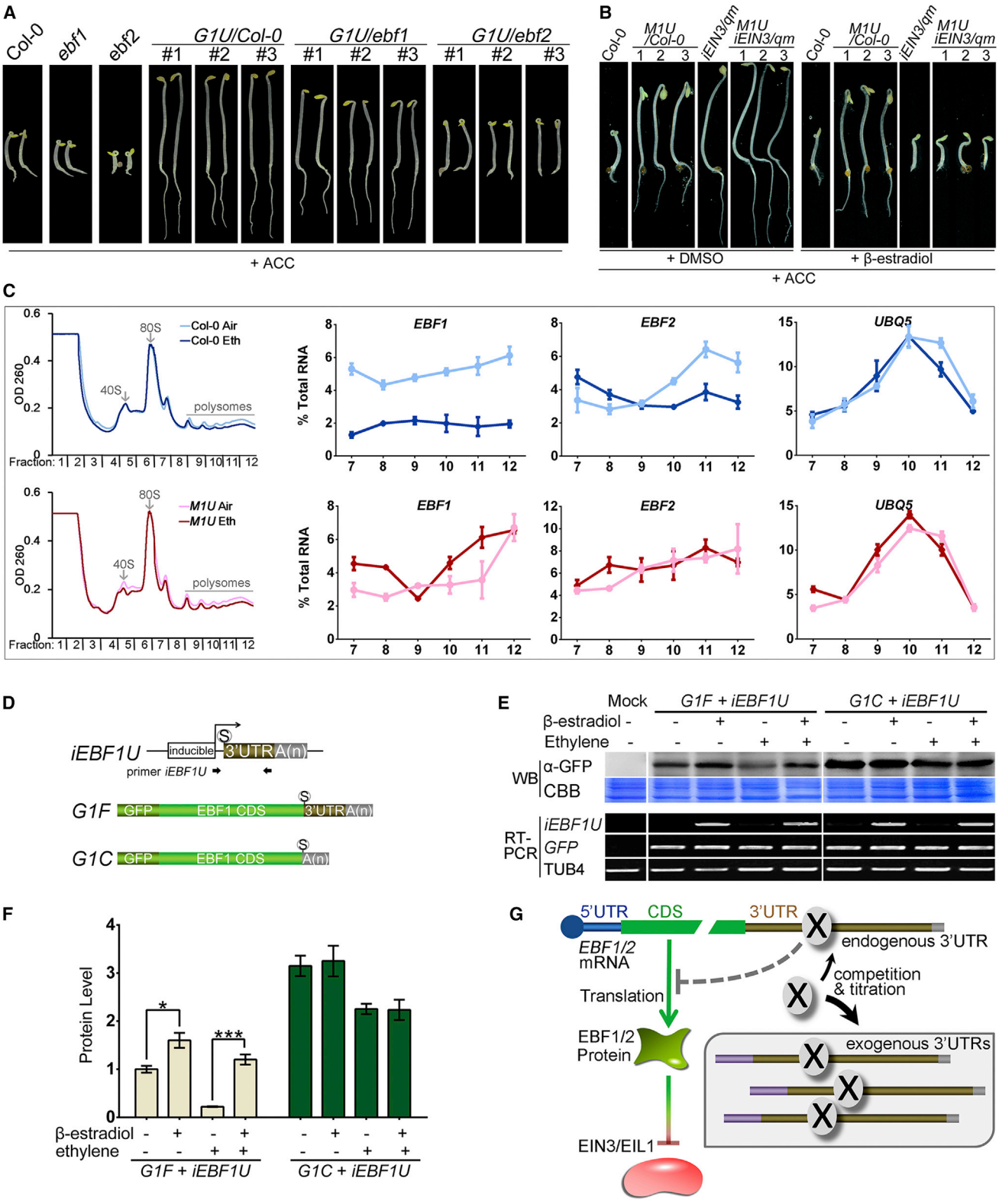


Figure 2. Overexpression of *EBF1 3′ UTR* Enhances the Translation of Endogenous *EBF1/2* mRNAs

(A and B) Triple response phenotypes of etiolated transgenic seedlings expressing *G1U* (*GFP-EBF1 3′ UTR*) treated with ACC (A), and seedlings expressing *M1U* (*MYC-EBF1 3′ UTR*) treated with ACC in combination with DMSO or β-estradiol (B). *iEIN3/qm* is the β-estradiol-induced *EIN3-Flag* in the *ein3 eil1 ebf1 ebf2* quadruple mutant background, which was used to substitute for the lethal *ebf1 ebf2* double mutant (An et al., 2010).

(legend continued on next page)

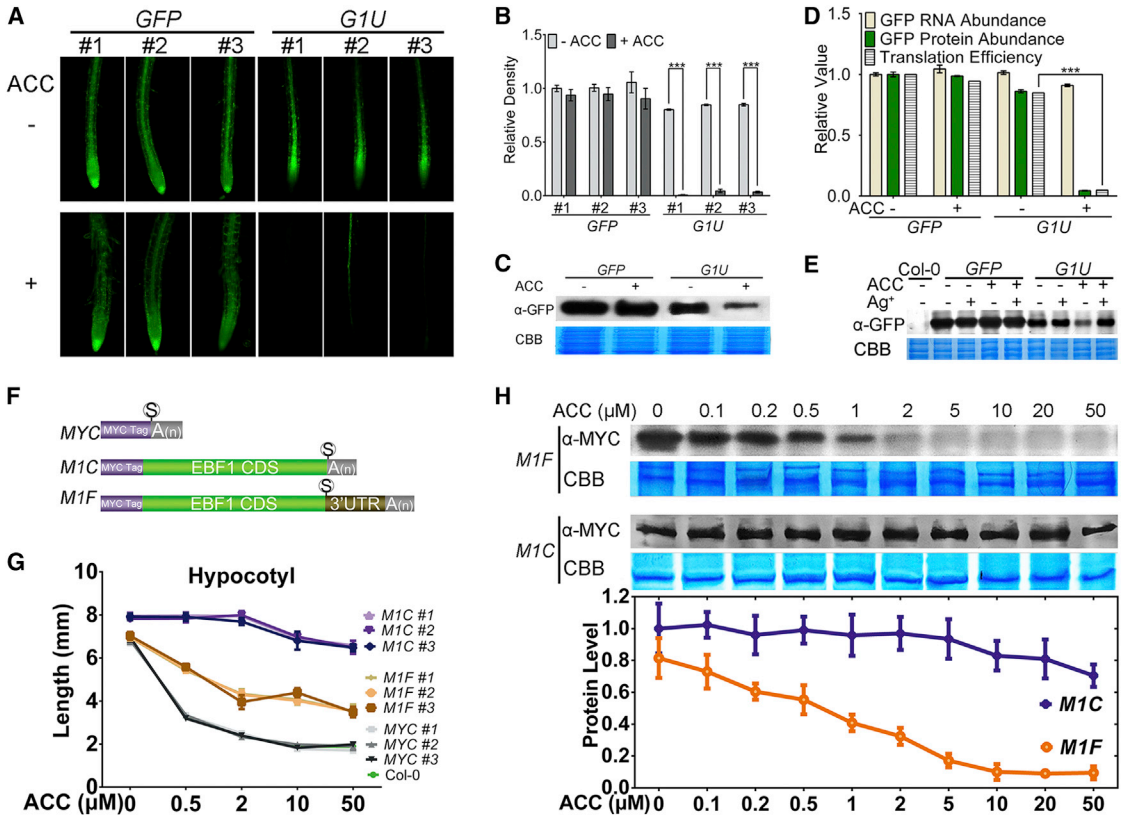


Figure 3. *EBF1* 3′ UTR Confers Translational Repression to Its Fusion Transcripts in Response to Ethylene

(A and B) GFP fluorescence in the roots of three independent transgenic seedlings expressing *GFP* or *G1U* (*GFP-EBF1 3′ UTR*) with (+) or without (–) ACC treatment (A) and the relative quantifications of GFP fluorescence (B). ***p < 0.001. Mean ± SD, n > 20 roots.

(C) Immunoblot assays showing GFP protein abundance in whole etiolated seedlings with (+) or without (–) ACC treatment.

(D) qRT-PCR analysis of *GFP* mRNAs and quantification of GFP proteins in (C). The ratio of protein to mRNA abundance was defined as the translation efficiency. ***p < 0.001; calculations based on three biological repeats.

(E) Immunoblot assays showing GFP protein abundance in etiolated seedlings treated with (+) or without (–) ACC and/or silver ion.

(F) Structures of *MYC*, *M1C* (*MYC-EBF1* CDS), and *M1F* (*MYC-EBF1* full length containing CDS and 3′ UTR) transcripts.

(G) Hypocotyl lengths of etiolated seedlings of three independent transgenic lines expressing *MYC*, *M1C*, and *M1F*. Mean ± SD, n > 20.

(H) Immunoblot assays indicating MYC-EBF1 protein abundances (top) and their relative quantifications (bottom) in seedlings treated with increasing doses of ACC. Calculations were based on three biological repeats.

See also Figure S3.

EIN2 Is Essential for 3′-UTR-Mediated Translational Repression of *EBF1* mRNA

We next investigated the role of key ethylene signaling components in 3′-UTR-mediated translational regulation. The ethylene-induced repression of *G1U* mRNA translation, manifested by reduced GFP fluorescence, was similarly observed in Col-0 and *ein3 eil1*, but not in *ein2* and a receptor mutant *etr1* (Figures 4A and S4A), suggesting that the upstream signaling components including the receptors and EIN2 are required for 3′-UTR-mediated translational repression, whereas EIN3/EIL1 are not. Expression of a β-estradiol-inducible version of *EIN2* was sufficient to restore such translation inhibition in *ein2*, and

(C) Polysome profiling assays with sucrose density gradient accompanied by qRT-PCR to analyze translational status of *EBF1/2* mRNAs. A_{254} absorption was monitored together with fractionation (left). The fractions containing 40S, 80S of ribosome, and polysomes are indicated. The abundance of *EBF1* and *EBF2* mRNA in each fraction was detected by qRT-PCR and quantified as a percentage relative to their total amount (right). *UBQ5* mRNA was used as a reference.

(D) Structures of *iEBF1U* (β-estradiol-inducible *EBF1 3′ UTR*) transcript, *G1F* (*GFP-EBF1* full length containing CDS and 3′ UTR) and *G1C* (*GFP-EBF1* CDS). Arrows indicate the primer pair used to analyze the expression of *iEBF1U*.

(E) Coexpression of *G1F* or *G1C* together with *iEBF1U* in etiolated seedlings treated with or without ethylene and β-estradiol for 4 hr before RT-PCR and western blotting analysis. Protein loading was manifested by Coomassie brilliant blue (CBB) staining.

(F) Quantitative measurements of GFP-EBF1 proteins in (E) based on three biological repeats. *p < 0.05; ***p < 0.001.

(G) A translational interference model proposes that the exogenously overexpressed *3′ UTRs* enhance the translation of endogenous *EBF1/2* mRNAs by competing with their inherent 3′ UTRs and thus titrating unknown repressor X bound to 3′ UTRs.

See also Figure S2.

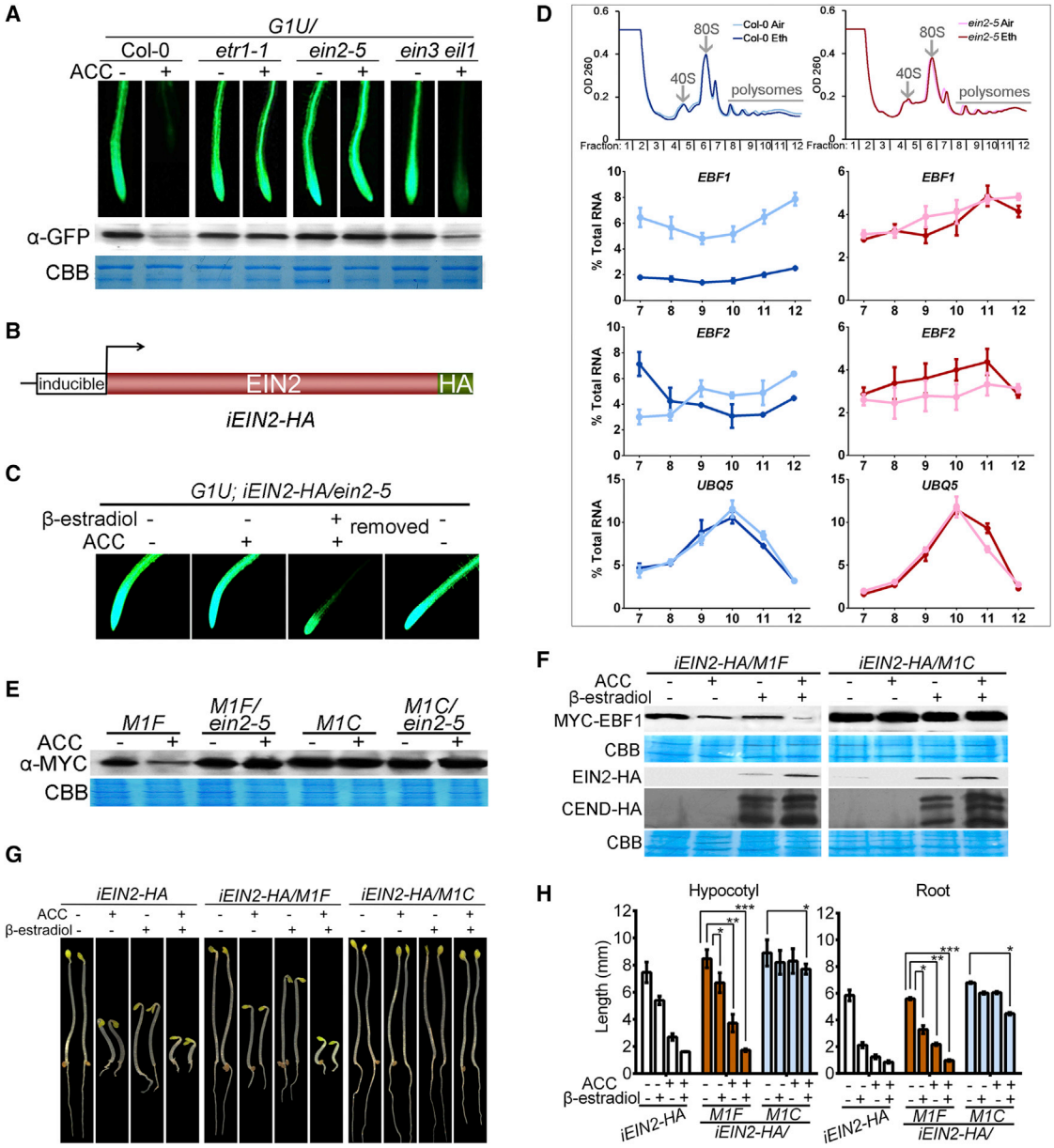


Figure 4. EIN2 Is Required for *EBF1* 3′-UTR-Mediated Translational Repression

(A) GFP fluorescence in the roots of etiolated seedlings expressing *G1U* (*GFP-EBF1 3′ UTR*) in different genotype backgrounds (top). Immunoblot assays showing GFP protein abundance in whole seedlings (bottom).

(B) Structure of the β-estradiol-inducible *EIN2-HA* gene (*iEIN2-HA*).

(C) GFP fluorescence in the roots of etiolated seedlings transiently treated with or without ACC and β-estradiol for 6 hr. "Removed," removal of both ACC and β-estradiol.

(D) Profiles of polysome-associated *EBF1*, *EBF2*, and *UBQ5* mRNAs in Col-0 and *ein2-5*.

(E) Immunoblot assays showing MYC-EBF1 protein abundance in etiolated seedlings of transgenic plants expressing *M1F* (*MYC-EBF1 CDS+3′UTR*) or *M1C* (*MYC-EBF1 CDS*).

(F) Immunoblot assays showing MYC-EBF1 and EIN2-HA protein abundance in transgenic plants expressing *iEIN2-HA* together with *M1F* or *M1C*. Note that multiple processed C-terminal fragments of induced EIN2-HA (CEND-HA) were also shown.

(G) Triple response phenotypes of etiolated seedlings corresponding to (F).

(H) Quantitative measurements of hypocotyls (left) and roots (right) of etiolated seedlings in (G). *p < 0.05; **p < 0.01; ***p < 0.001. Mean ± SD, n > 20.

See also Figure S4.

the removal of β-estradiol led to the efficient translation of *G1U* again (Figures 4B and 4C). A similar scenario was observed with transiently expressed β-estradiol-inducible *EIN2* and *G1U* in tobacco (Figure S4B), supporting that EIN2 is essential for *EBF1* 3′ UTR-directed translational repression.

To gain further evidence for EIN2-regulated *EBF1/2* mRNA translation, we compared the polysome profiles of *EBF1/2* mRNAs between Col-0 and *ein2* (Figure 4D). The polysome profiles of *EBF1/2* mRNAs remained virtually unchanged in *ein2* when treated with ethylene, in contrast with the apparent ethylene-induced polysome profile shifts observed in Col-0 (Figures 2C and 4D). Meanwhile, we found that the ethylene-evoked translational repression of *M1F* (*MYC-EBF1* full-length transcript) was abolished in *ein2* (Figures 4E and S4C), but exacerbated by addition of *EIN2* function (Figure 4F). By contrast, the translation of *M1C* (*MYC-EBF1 CDS*) remained unaffected upon depletion or addition of *EIN2* (Figures 4E, 4F, and S4C). Furthermore, the partial ethylene-insensitivity phenotype of *M1F* transgenic plants was largely suppressed by the overexpression of *EIN2*, whereas the strong ethylene insensitivity of *M1C* was hardly affected (Figures 4G and 4H). Taken together, these results indicate that 3′ UTR is a critical ethylene-responsive element to repress *EBF1* translation, and EIN2 is necessary and sufficient for directing such translational repression.

EIN2-Directed Translational Repression Is Mediated by PolyU Motifs of *EBF1/2* 3′ UTRs

We next dissected the functional *cis* elements within the *EBF1/2* 3′ UTRs by utilizing a dual-construct translation analysis system in tobacco leaves, in which a *3′ UTR* fragment of interest was fused with the *GFP* coding region, together with *mCherry* as the internal control in the same reporter construct (Figures 5A and S5A). The GFP intensities relative to mCherry intensities were calculated to indicate the translation efficiency of *GFP* mRNA (Figure 5B). Whereas the translation of *GFP* alone was not altered by introduction of *EIN2* and/or ACC application, the translation of *G1U* and *G2U* (*GFP* fused with *EBF1/2* 3′ UTR, respectively) was remarkably repressed by either expression of *EIN2* or ACC application (Figures S5B–S5D and S5K), and to a further extent when combining these two treatments (Figure S5C). As a control, expression of EIN3 protein had no effect on the translation of *G1U* (Figures S5E–S5H). These results confirmed the inhibitory effect of *EBF1/2* 3′ UTRs on translation in an EIN2-dependent manner.

EBF1 3′ UTR was arbitrarily segmented into five fragments ranging from 98 to 150 nt in length (Figure 5C). Three fragments, including *1Ua*, *1Ub* and *1Ud*, were able to mediate EIN2-induced translational repression (Figure 5D). Using the computation algorithm MEME and RNAfold, we identified a total of 7 poly-uridylates motifs in the predicted stem-loop structure within these three fragments (Figure 5E). These sequences were designated as Ethylene Responsive RNA elements containing Poly-Uridylates (*ERR-PolyU*, or *EPU* for short) (Figure 5E). Deletion of *EPUs* in each fragment or all seven *EPUs* in *1U,* which did not change their overall predicted secondary structures (Figure S5I), eliminated EIN2-directed translational repression (Figure 5F). Similarly, five *EPUs* were found in *EBF2* 3′ UTR (Figure S5J), and they were all required for *2U* to mediate EIN2-induced translational inhibition (Figure S5K). Sequence alignment of *EBF* 3′ UTRs from different plant species revealed that PolyU motifs are among the most conserved regions (Figures S5L and S5M), suggesting the 3′-UTR-mediated translational regulation as a well-preserved mechanism of ethylene signaling.

To further investigate the role of *EPUs* in relaying ethylene signaling, we generated the transgenic plants expressing either the *GFP-EBF1* full-length transcript driven by its own promoter (*pEBF1::G1F*) or seven *EPUs*-depleted version (*pEBF1::G1FΔ7U*) in *ebf1* mutant background. While expression of *pEBF1::G1F* rescued *ebf1* to the wild-type level, the *pEBF1::G1FΔ7U/ebf1* seedlings exhibited nearly complete ethylene insensitivity, phenocopying *pEBF1::G1C/ebf1* plants (*GFP-EBF1 CDS* driven by its own promoter) (Figure 5G). Consistent with the ethylene-response phenotype, the levels of EIN3 protein were much lower in both *pEBF1::G1FΔ7U/ebf1* and *pEBF1::G1C/ebf1* than that in Col-0 or *pEBF1::G1F/ebf1*, whereas the GFP-EBF1 protein was more abundant in the former two lines, particularly under ethylene treatment (Figure S5N). These results suggest that *EPU*-mediated translational inhibition plays a key part in regulating EBF1 protein abundance as well as ethylene signal transduction.

From *1Ud*, we selected a region harboring two *EPUs* that is predicted to form a hairpin structure (Figure 5H), and repeated it three times to construct an artificial *3′ UTR* that possessed six *EPUs* (*6x EPU*) (Figure 5H). Similar to *G1U*, the translation of *GFP-6x EPU* mRNA was highly reduced upon *EIN2* induction (Figure 5I). Furthermore, transgenic overexpression of *GFP-6x EPU* but not *GFP-1UΔ7U* conferred ethylene insensitivity phenotype (Figure S5O). Together, these results demonstrate that *EPUs* mediate the EIN2-directed translational repression of *EBF1/2*, which represents a crucial mechanism of ethylene signaling.

We also examined the functional domain of EIN2 in translational repression. By taking advantage of the tobacco system, we narrowed down the C-terminal end of EIN2 fragments (CEND) to amino acids (aa) 654–1272 that were required for translational repression (Figures 5J, 5K, and S5P). Within this region, a predicted nuclear localization signal (NLS, aa 1262–1269, LKRYKRRL) was previously identified to be required for the nuclear translocation as well as the functionality of CEND (Ju et al., 2012; Qiao et al., 2012; Wen et al., 2012). We found that deletion or mutation of this NLS region also disrupted the function of CEND in translational repression (Figures 5J and 5K). Interestingly, replacement of the NLS with a distinct K/R-rich NLS sequence (NLS': KPKKKRKV) was able to relocate CEND into the nucleus but failed to restore its translational repression ability (Figures 5K and S6G). Together, these results suggested that the short motif (aa 1262–1269) was also critical for the translational repression function of EIN2 independent of its being a nuclear localization signal.

Association and Co-localization of EIN2 with *EBF1 3′ UTR* in Cytoplasmic Foci

We next investigated how EIN2 imposes translational repression of *1U/2U*-containing mRNAs. We first examined whether EIN2 associates with *1U in vivo*. RNA-immunoprecipitation assays (RNA-IP) in tobacco leaves indicated that EIN2 preferentially associated with mRNAs containing *1U* (*G1U, M1U*), but not

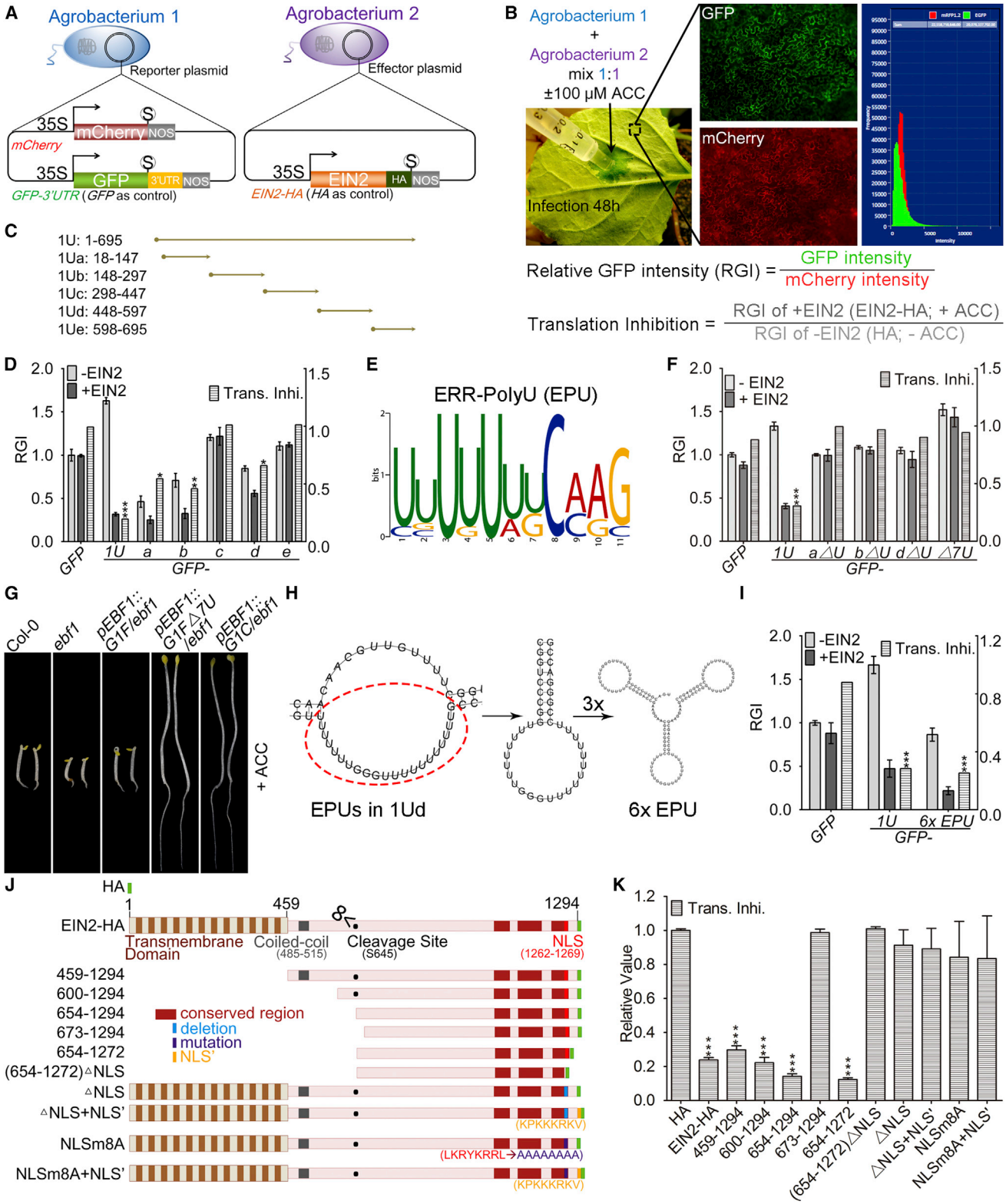


Figure 5. PolyU Motifs in *EBF1* 3′ UTR Are Necessary and Sufficient for EIN2-Directed Translational Inhibition

(A and B) Plasmids used in the dual-construct translation analysis system (A) as well as the workflow (B). The reporter plasmid harbors the reference gene *mCherry* and the reporter gene *GFP-3′UTR* (*GFP* as control). The effector plasmid possesses *EIN2-HA* (*HA* as control) (A). ACC application was used to further activate the

(legend continued on next page)

with *GFP* mRNA alone, and ethylene enhanced the association between EIN2 and *G1U* mRNA (Figures 6A and S6A).

Next, we sought to examine the subcellular localization and dynamics of *1U*-containing mRNAs and EIN2. We adopted the MS2 system (Bertrand et al., 1998) to directly visualize the subcellular localization of *1U*-containing mRNAs. In this system, YFP was fused to the C terminus of MS2 coat protein (MY), and six tandem repeats of MS2 binding sites (6X MS2bs) were inserted into *M1U* to produce a reporter RNA *MYC-6X MS2bs-1U* (*M6U*), while a reporter RNA *MYC-6X MS2bs* (*M6*) served as a negative control (Figures 6B, S6B, and S6C). RNA-IP assay revealed the association of EIN2 and *M6U in vivo* (Figure S6A), and transgenic plants overexpressing *M6U* showed ethylene-insensitivity phenotypes (Figure S6D), demonstrating the functionality of this fusion RNA. In the absence of ethylene, *M6U* was observed to spread in the cytoplasm and concentrate in the nucleus, similar to the distribution pattern of *M6* (Figure 6C). Notably, ethylene treatment specifically induced *M6U* but not *M6* to form granules in the cytoplasm (Figures 6C and 6E).

Meanwhile, we found that ethylene treatment can also induce a proportion of EIN2 to form cytoplasmic foci in addition to its nuclear accumulation (Figures S6E and S6F; Movies S1 and S2). In the presence of ethylene, a portion of EIN2 protein and *M6U* mRNA were co-localized in cytoplasmic foci (Figure 6D). Furthermore, the cytoplasmic foci formation of *M6U* was abolished in *ein2* (Figure 6E), suggesting the requirement of EIN2 for foci formation of *1U*-containing mRNA. Taken together, these results suggest that ethylene promotes the association of EIN2 with *1U*, which in turn is targeted to cytoplasmic foci.

We further found that EIN2 CEND (aa 459–1294) as well as the minimal functional fragment of EIN2 (aa 654–1272) were also able to form cytoplasmic foci under ethylene treatment, whereas all the translation-dysfunctional fragments of EIN2, including aa 673–1294, deletion or mutation of NLS, failed to form foci in the cytoplasm (Figure S6G). It is noteworthy that the addition of another functional NLS sequence (NLS') could not restore the cytoplasmic foci formation of NLS-deleted or -mutated EIN2 (Figure S6G). Together with the observations made in Figure 5K, these results demonstrate that the NLS sequence of EIN2 is critical for its cytoplasmic foci formation as well as translational regulation function.

P-Body Is Involved in *EBF1/2* 3′-UTR-Mediated Translational Repression by EIN2

Given that *EBF1/2* RNAs are subjected to the regulation by EIN5, an exoribonuclease associated with processing body (P-body) (Decker and Parker, 2012; Xu and Chua, 2011), and that *1U*-containing mRNA forms cytoplasmic foci, we next determined whether *1U* directs its fusion mRNA to P-body. Upon ethylene treatment, both *M6U* mRNA and EIN2 protein were partly co-localized with EIN5 in cytoplasmic foci (Figures 7A and 7B), indicative of their P-body localization. Additionally, yeast-two-hybrid (Y2H) and luciferase complementation imaging (LCI) assay indicated that EIN2 CEND interacted with EIN5 (Figures 7C and 7D). Co-immunoprecipitation (Co-IP) assays revealed that EIN5 associated with EIN2 mainly in the presence of RNA, as treatment with RNase largely diminished EIN5-EIN2 association (Figure 7E). Furthermore, we found that several other P-body components, such as PAB2, PAB4 and PAB8 (Decker and Parker, 2012), also interacted with EIN2 CEND in yeast and plant cells in a RNA-dependent manner (Figures 7C, 7D, and S7A–S7D). In keeping with these biochemical results, knockout mutants of P-body component genes, such as *EIN5*, *PAB2*, *PAB8*, and *UPF1*, exhibited reduced ethylene sensitivity manifested by compromised triple response phenotypes, target gene expression, and EIN3 protein accumulation (Figures 7F and S7E–S7G). The combinations of these mutants led to increasing severity of ethylene-insensitivity phenotypes, particularly for the *ein5 upf1 pab2 pab8* quadruple mutant, which exhibited strong insensitivity to ethylene (Figures 7F and S7E). Therefore, several P-body components act cooperatively to repress the translation of mRNAs harboring *1U*.

Although UPF1 was not detected to physically interact with EIN2, we observed the binding of UPF1 to *1U* (Figure S7H), consistent with its function as a non-selective RNA binding protein (Hogg and Goff, 2010). The comparable mRNA levels of *EBF1* and *EBF2* between the P-body mutants and wild-type plants further supported a control of translation rather than transcription or RNA decay of *EBF1/2* by P-body (Figure S7F). Taken together, we proposed that after activation by ethylene, EIN2 CEND associates with the 3′ UTR of *EBF1/2* mRNAs and targets them to P-body via interacting with multiple P-body components, thus repressing the translation of *EBF1* and *EBF2*, resulting in EIN3/EIL1 accumulation and ethylene responses (Figure 7G).

DISCUSSION

A Cytoplasmic Mode of EIN2 Action in Ethylene Signaling

Recently, three groups have uncovered a "cleave and shuttle" mode of EIN2 action, wherein its C-terminal end (CEND) is

EIN2-HA protein. Translational inhibition was calculated by relative GFP intensity in the presence of EIN2-HA and ACC application (RGI of +EIN2) normalized with that without EIN2-HA and ACC (RGI of -EIN2) (B).

(C and D) Fragments of *1U* (*EBF1 3′ UTR*) and their effects on the translation of *GFP* mRNA with or without EIN2 function. *p < 0.05, **p < 0.01, ***p < 0.001. Calculations of translational inhibition were based on three biological repeats and the value of *GFP* control was set as 1.

(E) The PolyU ethylene responsive RNA elements (termed as *ERR-PolyU* or *EPU*) shared in the fragments *1Ua*, *1Ub*, and *1Ud*.

(F) The effects of *EPUs* on the translational inhibition. *ΔU*, deletion of *EPU*. *Δ7U*, deletion of all seven *EPUs* in full-length *1U*. ***p < 0.001.

(G) Triple response phenotype of etiolated seedlings in the presence of ACC. *G1F* (*GFP-EBF1* full length containing CDS and 3′ UTR), *G1FΔ7U* (*G1F* with all seven *EPUs* deleted), and *G1C* (*GFP-EBF1* CDS) were all driven by native *EBF1* promoter (*pEBF1*) and expressing in *ebf1* mutant.

(H and I) An engineered *6x EPU* fragment and its effect on translational inhibition upon EIN2 activation. G and C bases were added to the two *EPUs* in *1Ud* to produce a stem-loop structure, which was repeated three times to generate *6x EPU*.

(J and K) Scheme for different EIN2 fragments and their inhibitory effect on *G1U* (*GFP-EBF1 3′ UTR*) translation. ΔNLS indicates the deletion of the predicted nuclear localization signal. NLS′ represents a distinct NLS sequence. NLSm8A means the substitution of the NLS motif with eight alanine residues. ***p < 0.001. See also Figure S5.

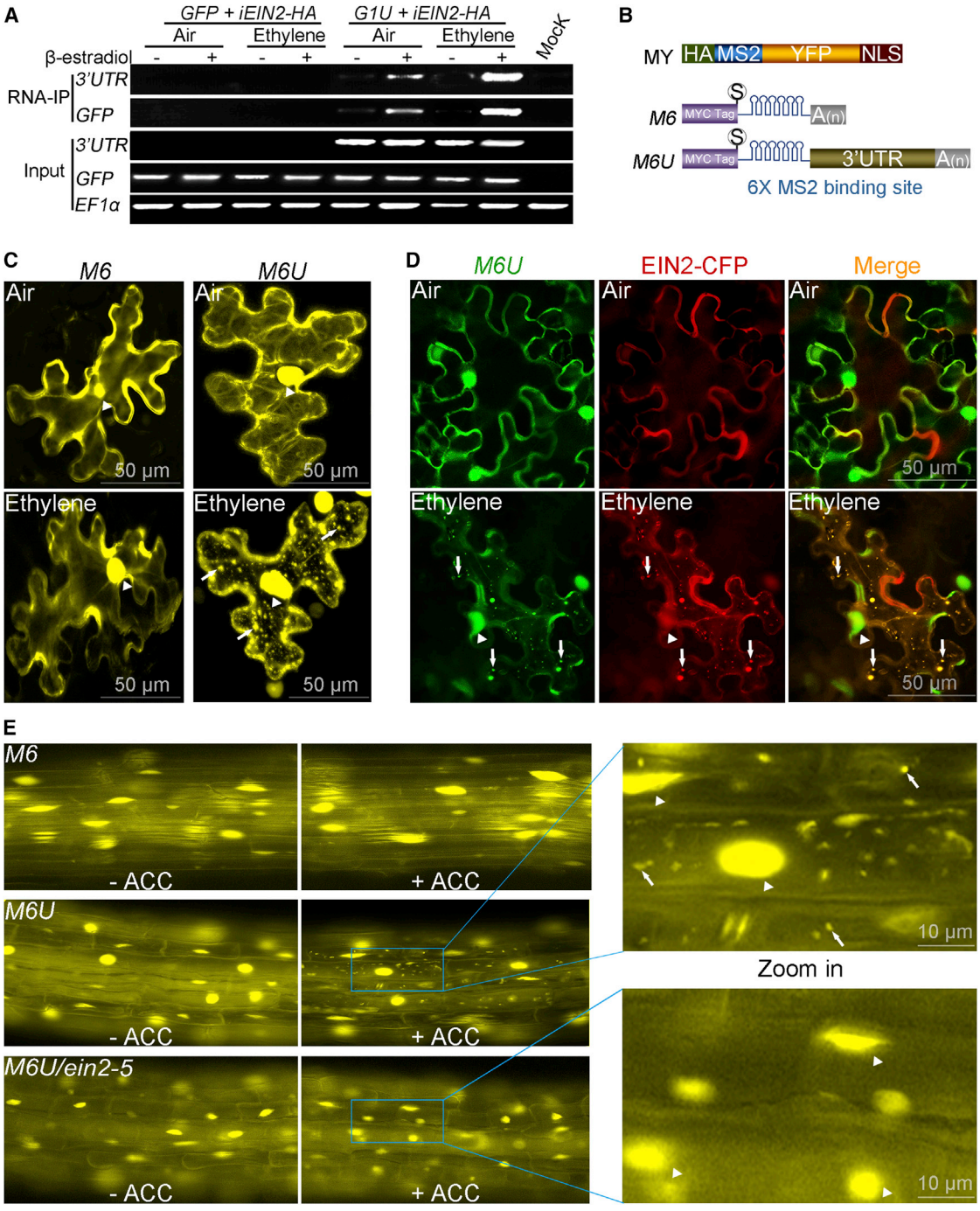


Figure 6. Ethylene Induces the Association and Co-localization of *EBF1 3′ UTR* with EIN2 in Cytoplasmic Foci

(A) RNA-IP assays indicating the association between EIN2 and *G1U* (*GFP-EBF1 3′ UTR*) in tobacco leaves. *GFP* acts as a negative control. (*iEIN2-HA*: β-estradiol-inducible *EIN2-HA*).

(B) Schematic diagrams of the MS2/RNA-MS2bs system. MY means MS2 coat protein linked with YFP; *M6* and *M6U*, *MYC-6X MS2 binding site* and *MYC-6X MS2 binding site -EBF1 3′ UTR*, respectively. S in a circle, stop codon.

(C) YFP fluorescence revealing the subcellular localization of *M6* and *M6U* RNAs in tobacco leaves treated with or without ethylene. Arrows mark cytoplasmic foci, while triangles indicate nuclei.

(D) Co-localization of EIN2-CFP and *M6U* in tobacco leaves upon ethylene treatment.

(E) The subcellular localization of *M6* or *M6U* in transgenic *Arabidopsis* seedlings treated with or without ACC. Right panels are zoom-in images of the boxed areas in left.

See also Figure S6.

processed and translocated into the nucleus to activate ethylene signaling (Ju et al., 2012; Qiao et al., 2012; Wen et al., 2012). Here, we report another mechanism of EIN2-mediated ethylene signaling, whereby EIN2 imposes translational repression of *EBF1/2* mRNA in cytoplasmic P-body compartments. This cytoplasmic mode of EIN2 action was revealed by several lines of evidence: (1) EIN2 and ethylene treatment inhibit the translation of *EBF1/2* mRNAs. (2) EIN2 is both necessary and sufficient for the translational repression of *1U*-containing mRNAs. (3) EIN2 is colocalized and associated with *1U*. (4) *1U*, EIN2 and the EIN5 are co-localized in P-bodies upon ethylene treatment. (5) EIN2 interacts with several P-body factors including EIN5, PAB2/4/8. (6) Mutations in P-body protein genes led to evident ethylene insensitivity, particularly in combinations. Together with previous studies, our discovery illustrates that EIN2 guarantees the accumulation of key transcription factors EIN3/EIL1 in response to ethylene through at least two parallel mechanisms (Figure 7G). The cytoplasmic function of EIN2 is critical for quickly shutting down the protein synthesis of *EBF1/2*, leading to rapid depletion of EBF1/2 proteins due to its proteasomal degradation (An et al., 2010). Meanwhile, a subset of CEND is translocated into the nucleus to further stabilize and/or activate EIN3/EIL1 directly or indirectly (Ji and Guo, 2013).

It has been previously reported that ethylene application causes polysome prevalence during the ripening of pear and avocado fruits, suggesting a positive regulation of translation by ethylene (Drouet and Hartmann, 1979; Tucker and Laties, 1984). In an accompanying study, Merchante et al. (2015) used a plant-optimized genome-wide ribosome footprinting technique and successfully identified a group of mRNA targets that are upregulated or downregulated by ethylene at translational level. Of these targets, *EBF1* and *EBF2* are prominent as ethylene-repressed mRNAs that are dependent on EIN2 but not EIN3/EIL1, as observed also in our study. Thus, the EIN2-dictated translational control represents an early signaling event that operates in the cytoplasm either in parallel with or prior to the nuclear signaling cascade. Interestingly, this research together with previous studies (Qiao et al., 2012; Wen et al., 2012) revealed that a predicted NLS motif in the very C terminus of EIN2 is essential for its functions in both cytoplasm and nucleus. Given the recent finding that NLS is also critical for the association between EIN2 and the ethylene receptor ETR1 on ER (Bisson and Groth, 2015), it remains to be addressed how such short motif is involved in seemingly distinctive subcellular signaling events.

EBF1/2 3′ UTRs Function as Critical Ethylene-Responsive and Signal-Relaying Elements

In mammals, 3′ UTRs targeted by microRNAs are critical for the regulation of proto-oncogenes and tumorgenesis (Mayr and Bartel, 2009). Recent efforts were taken to systematically analyze human 3′ UTRs, and dozens of novel *cis*-regulatory elements were identified that affect mRNA stability and translation (Oikonomou et al., 2014; Zhao et al., 2014). Our study revealed that the 3′-UTR-mediated translational repression of *EBF1/2* is vital for relaying the ethylene signal in plants. The biological significance of this repression was demonstrated by the findings that deletion of *EBF1* 3′ UTR or the *EPU* motifs greatly enhanced the translation of *EBF1* mRNA and led to nearly complete ethylene insensitivity (Figures 4G and 5G). We further identified multiple PolyU motifs in the loop of predicted stem-loop structures (*EPUs*) as functional *cis* elements shared in *EBF1* and *EBF2* 3′ UTRs (Figure 5). Considering the conservation of EIN2 from green algae to land plants (Ju et al., 2015), and of PolyU motifs in the *EBF* 3′ UTR sequences from different plant species (Figures S5L and S5M), we believe that the 3′-UTR-mediated translational regulation might be an evolutionarily widespread mechanism of ethylene signaling.

Furthermore, our study indicates that the ethylene-induced *EBF1/2* translational repression is likely to be achieved by targeting *EBF1/2* transcripts into P-body in an EIN2-dependent manner. Although our initial in vitro pull-down assays failed to detect their direct binding, RNA IP experiment revealed the association of EIN2 with *EBF1* 3′ UTR in vivo (Figure 6A). It raises the possibility that some unidentified RNA binding proteins, which could specifically recognize PolyU motifs of 3′ UTRs, directly or indirectly interact with EIN2 and tether it to *EBF1/2* mRNAs (Figure 7G). EIN2, therefore, may act as a hormone-activated switch to target *EBF1/2* mRNAs (and probably other mRNAs as well) to P-bodies via interaction with P-body proteins.

Cytoplasmic foci, including P-body and stress granules, have been observed in plant cells under myriad stress conditions (Maldonado-Bonilla, 2014). The importance of P-body in ethylene signaling was manifested as mutants of several P-body components led to reduced ethylene sensitivity (Figure 7F). Therefore, ethylene, well known as a stress hormone, might adopt the translational repression mechanism via P-body to quickly shut down gene expression under adverse stress conditions.

Utilizing the Translational Interference Effect of 3′ UTR to Modulate Gene Function

Overexpression of the coding sequence of a gene had been widely utilized as a powerful genetic tool to study the gene functions in animals and plants (Prelich, 2012). In this study, we demonstrated that overexpression of *3′ UTR* could also result in remarkable interference with the function of their cognate genes as well as the signaling output. Several lines of evidence supported that the exogenous expression of *3′ UTR* leads to the enhancement or de-repression of the endogenous *EBF1/2* mRNA harboring the same or related 3′ UTR in a *trans*-acting manner. Given the strong phenotype of *3′-UTR*-overexpressing transgenic plants, our study offers an alternative tool to study and regulate the function of genes in vivo.

The translational interference effect of 3′ UTR illustrated in this work is reminiscent of the action of microRNA sponges (Ebert et al., 2007) as well as competitive endogenous RNA (*ceRNA*) in mammals (Denzler et al., 2014; Salmena et al., 2011), and microRNA target mimics in plants (Franco-Zorrilla et al., 2007), all of which share a common underlying mechanism referred to as molecular titration (Bosson et al., 2014; Buchler and Louis, 2008). As such, the accumulation of *3′ UTR* fragments, as observed in *ein5* (Souret et al., 2004), might hold biological importance, such as to coordinate or buffer the translational regulation of related mRNAs. In the future, a more systematic identification and study of 3′ UTRs in plants and animals would

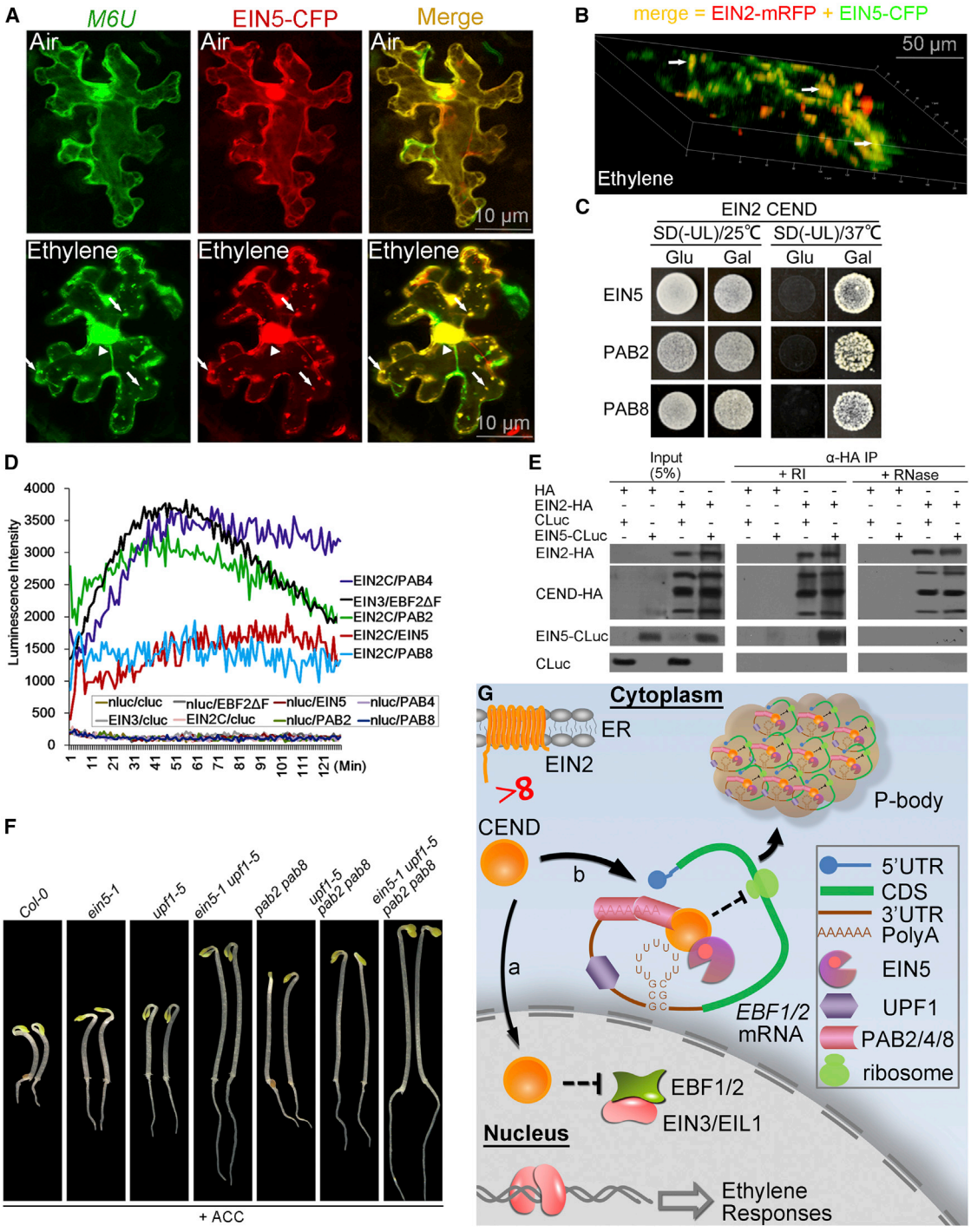


Figure 7. EIN2 Co-localizes with *EBF1 3′ UTR* in P-Body and Interacts with Multiple P-Body Components

(A) Co-localization of *M6U* (*MYC-6X MS2 binding site-EBF1 3′ UTR*) (green) and EIN5 (red) in P-bodies of tobacco leaves. Arrows mark cytoplasmic foci (P-bodies), while triangles indicate nuclei.

(B) A 3D image showing partial co-localization of EIN2 (red) and EIN5 (green) in P-bodies (arrow).

(C) Yeast-2-hybrid assays indicating the interactions between EIN2 CEND (889–1294) and EIN5 as well as PAB2/8.

(D) Luciferase complementation imaging (LCI) assays manifesting the interaction between EIN2 CEND and P-body components in *Arabidopsis* protoplasts. Combinations in the right list show strong interaction, while the others in the bottom box are either negative controls or exhibit no interaction.

(E) Co-immunoprecipitation assays indicating the association between EIN2 and EIN5 in the presence of RNA. Immunoblot assays showing the amount of expressed proteins in tobacco leaf extracts (input) and after IP with anti-HA antibody. HA and CLuc were used as negative controls. RI, RNase inhibitor.

(legend continued on next page)

provide new information about gene functions as well as their regulatory mechanisms.

EXPERIMENTAL PROCEDURES

Arabidopsis Materials and Growth Conditions

The ecotype Columbia (Col-0) was the parent line for all mutants and transgenic plants used in this study. Transgenic lines in different genetic backgrounds were constructed by genetic crosses. Unless otherwise stated, all *Arabidopsis* seedlings were grown on MS medium supplied with or without 10 μM ACC, or other chemicals, for 3–4 days. For transient treatments, 100 μM ACC or 10 ppm ethylene was used for seedlings, and 100 μM ACC or 100 ppm ethylene was used for tobacco leaves.

Polysome Profiling

Arabidopsis polysomes were fractionated over sucrose gradients as described (Missra and von Arnim, 2014) with minor modifications. 3-day-old etiolated seedlings were treated with 10 ppm ethylene for 4 hr and then ground in liquid nitrogen followed by resuspension in polysome extraction buffer. Supernatant was loaded onto a 15%–60% sucrose gradient and spun in a Beckman SW40Ti rotor at 40,000 rpm for 4 hr at 4°C. We collected 12 fractions by a gradient fractionator. Total RNA in each fraction was isolated using TRIZOL reagent (Life Technologies) and then subjected to reverse transcription and real-time PCR analysis.

RNA Immunoprecipitation

4-week-old tobacco leaves were infected with the mixture of two agrobacterium strains. Two days after agroinfiltration, the tobacco leaves were treated with air or ethylene for 4 hr and subsequently collected to be ground in liquid nitrogen, and protein/RNA complexes were extracted using two volumes of IP buffer. After removal of insoluble debris by centrifugation, cell extracts were incubated with anti-HA antibody (Sigma) for 2 hr on ice with occasional gentle mixing. The anti-HA-decorated extracts were incubated with pre-washed protein G agarose beads. The co-immunoprecipitated RNA was isolated by TRIZOL reagent (Life Technologies) and analyzed by qRT-PCR.

Co-Immunoprecipitation

One-month-old tobacco leaves were infected with the mixture of three agrobacterium strains. Protein samples prepared from tobacco leaves 48 hr after Agrobacterium-mediated infiltration were homogenized in ice-cold IP buffer with the volume ratio of 1/2. After centrifugation, lysates were supplemented extemporaneously with RNase inhibitor (Promega) or RNase (Promega) and then incubated for 2 hr at 4°C under gentle agitation in the presence of EZview anti-HA affinity gel (Sigma). Antibody-coupled agarose beads were washed and subsequently denatured to detect the IPed proteins using western blot.

See Supplemental Experimental Procedures for details on the above-described materials and methods, as well as additional methods and procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, one table, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.037.

AUTHOR CONTRIBUTIONS

W.L., M.M., and H.G. designed all experiments, analyzed data, and wrote the manuscript. W.L. and M.M. performed most of the experiments and prepared data. W.L. and Y.F. conducted the polysome profiling assays. W.L. and Y.W. performed microscopy. H.L., M.L., and F.A. contributed to the generation and analysis of myriad transgenic plants. M.M. and Y.M. conducted LCI, yeast, and tobacco dual-construct translational assays. Y.F. and M.M. helped prepare the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Robert H. Singer, Dr. Jianru Zuo, Dr. Jinsong Zhang, and Dr. Jianmin Zhou for their gift of vectors. We thank H.G.'s lab members for their technical assistance and critical discussions. This work was supported by the National Natural Science Foundation of China (91217305 and 91017010) and the National Basic Research Program of China (973 Program; 2012CB910902) to H.G. This study is also supported by the 111 project of Peking University.

Received: April 17, 2015
Revised: August 4, 2015
Accepted: August 31, 2015
Published: October 22, 2015

REFERENCES

Alonso, J.M., Hirayama, T., Roman, G., Nourizadeh, S., and Ecker, J.R. (1999). EIN2, a bifunctional transducer of ethylene and stress responses in *Arabidopsis*. Science *284*, 2148–2152.

An, F., Zhao, Q., Ji, Y., Li, W., Jiang, Z., Yu, X., Zhang, C., Han, Y., He, W., Liu, Y., et al. (2010). Ethylene-induced stabilization of ETHYLENE INSENSITIVE3 and EIN3-LIKE1 is mediated by proteasomal degradation of EIN3 binding F-box 1 and 2 that requires EIN2 in *Arabidopsis*. Plant Cell *22*, 2384–2401.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998). Localization of *ASH1* mRNA particles in living yeast. Mol. Cell *2*, 437–445.

Bisson, M.M., and Groth, G. (2011). New paradigm in ethylene signaling: EIN2, the central regulator of the signaling pathway, interacts directly with the upstream receptors. Plant Signal. Behav. *6*, 164–166.

Bisson, M.M., and Groth, G. (2015). Targeting plant ethylene responses by controlling essential protein-protein interactions in the ethylene pathway. Mol. Plant *8*, 1165–1174.

Bleecker, A.B., Estelle, M.A., Somerville, C., and Kende, H. (1988). Insensitivity to ethylene conferred by a dominant mutation in *Arabidopsis thaliana*. Science *241*, 1086–1089.

Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential *ceRNA* competition. Mol. Cell *56*, 347–359.

Buchler, N.E., and Louis, M. (2008). Molecular titration and ultrasensitivity in regulatory networks. J. Mol. Biol. *384*, 1106–1119.

Chang, C., and Stadler, R. (2001). Ethylene hormone receptor action in *Arabidopsis*. BioEssays *23*, 619–627.

Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.S., Schmitz, R.J., Urich, M.A., Kuo, D., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in *Arabidopsis*. eLife *2*, e00675.

Chao, Q., Rothenberg, M., Solano, R., Roman, G., Terzaghi, W., and Ecker, J.R. (1997). Activation of the ethylene gas response pathway in *Arabidopsis* by the nuclear protein ETHYLENE-INSENSITIVE3 and related proteins. Cell *89*, 1133–1144.

(F) Triple response phenotypes of etiolated seedlings of indicated genotypes.
(G) A schematic model depicting two branches of ethylene signaling pathway relaying from EIN2 on the ER membrane into nucleus to regulate EIN3/EIL1 protein stability. The cytoplasmic mode of EIN2 action revealed in this study is highlighted as the formation of P-bodies containing CEND, *EBF1/2* 3′ UTRs, and several P-body proteins including EIN5, PABs, and UPF1.
See also Figure S7.

Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harb. Perspect. Biol. *4*, a012286.

Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the *ceRNA* hypothesis with quantitative measurements of miRNA and target abundance. Mol. Cell *54*, 766–776.

Drouet, A., and Hartmann, C. (1979). Polyribosomes from pear fruit: changes during ripening and senescence. Plant Physiol. *64*, 1104–1108.

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods *4*, 721–726.

Ecker, J.R. (1995). The ethylene signal transduction pathway in plants. Science *268*, 667–675.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. *39*, 1033–1037.

Gao, Z., Chen, Y.F., Randlett, M.D., Zhao, X.C., Findell, J.L., Kieber, J.J., and Schaller, G.E. (2003). Localization of the Raf-like kinase CTR1 to the endoplasmic reticulum of *Arabidopsis* through participation in ethylene receptor signaling complexes. J. Biol. Chem. *278*, 34725–34732.

Guo, H., and Ecker, J.R. (2003). Plant responses to ethylene gas are mediated by SCF(EBF1/EBF2)-dependent proteolysis of EIN3 transcription factor. Cell *115*, 667–677.

Guo, H., and Ecker, J.R. (2004). The ethylene signaling pathway: new insights. Curr. Opin. Plant Biol. *7*, 40–49.

Hogg, J.R., and Goff, S.P. (2010). Upf1 senses 3'UTR length to potentiate mRNA decay. Cell *143*, 379–389.

Ji, Y., and Guo, H. (2013). From endoplasmic reticulum (ER) to nucleus: EIN2 bridges the gap in ethylene signaling. Mol. Plant *6*, 11–14.

Johnson, P.R., and Ecker, J.R. (1998). The ethylene gas signal transduction pathway: a molecular perspective. Annu. Rev. Genet. *32*, 227–254.

Ju, C., Yoon, G.M., Shemansky, J.M., Lin, D.Y., Ying, Z.I., Chang, J., Garrett, W.M., Kessenbrock, M., Groth, G., Tucker, M.L., et al. (2012). CTR1 phosphorylates the central regulator EIN2 to control ethylene hormone signaling from the ER membrane to the nucleus in *Arabidopsis*. Proc. Natl. Acad. Sci. USA *109*, 19486–19491.

Ju, C., Van de Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., and Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. Nature Plants *1*, 14004.

Kieber, J.J., Rothenberg, M., Roman, G., Feldmann, K.A., and Ecker, J.R. (1993). *CTR1*, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the raf family of protein kinases. Cell *72*, 427–441.

Konishi, M., and Yanagisawa, S. (2008). Two different mechanisms control ethylene sensitivity in *Arabidopsis* via the regulation of *EBF2* expression. Plant Signal. Behav. *3*, 749–751.

Maldonado-Bonilla, L.D. (2014). Composition and function of P bodies in *Arabidopsis thaliana*. Front. Plant Sci. *5*, 201.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

Merchante, C., Brumos, J., Yun, J., Hu, Q., Spencer, K.R., Enriquez, P., Binder, B.M., Heber, S., Stepanova, A.N., and Alonso, J.M. (2015). Gene-specific translation regulation mediated by the hormone-signaling molecule EIN2. Cell *163*, this issue, 684–697.

Missra, A., and von Arnim, A.G. (2014). Analysis of mRNA translation states in *Arabidopsis* over the diurnal cycle by polysome microarray. Methods Mol. Biol. *1158*, 157–174.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell Rep. *7*, 281–292.

Olmedo, G., Guo, H., Gregory, B.D., Nourizadeh, S.D., Aguilar-Henonin, L., Li, H., An, F., Guzman, P., and Ecker, J.R. (2006). ETHYLENE-INSENSITIVE5 encodes a 5′–>3′ exoribonuclease required for regulation of the EIN3-targeting F-box proteins EBF1/2. Proc. Natl. Acad. Sci. USA *103*, 13286–13293.

Potuschak, T., Lechner, E., Parmentier, Y., Yanagisawa, S., Grava, S., Koncz, C., and Genschik, P. (2003). EIN3-dependent regulation of plant ethylene hormone signaling by two *arabidopsis* F box proteins: EBF1 and EBF2. Cell *115*, 679–689.

Potuschak, T., Vansiri, A., Binder, B.M., Lechner, E., Vierstra, R.D., and Genschik, P. (2006). The exoribonuclease XRN4 is a component of the ethylene response pathway in *Arabidopsis*. Plant Cell *18*, 3047–3057.

Prelich, G. (2012). Gene overexpression: uses, mechanisms, and interpretation. Genetics *190*, 841–854.

Qiao, H., Shen, Z., Huang, S.S., Schmitz, R.J., Urich, M.A., Briggs, S.P., and Ecker, J.R. (2012). Processing and subcellular trafficking of ER-tethered EIN2 control response to ethylene gas. Science *338*, 390–393.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A *ceRNA* hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Souret, F.F., Kastenmayer, J.P., and Green, P.J. (2004). AtXRN4 degrades mRNA in *Arabidopsis* and its substrates include selected miRNA targets. Mol. Cell *15*, 173–183.

Tucker, M.L., and Laties, G.G. (1984). Interrelationship of gene expression, polysome prevalence, and respiration during ripening of ethylene and/or cyanide-treated avocado fruit. Plant Physiol. *74*, 307–315.

Wen, X., Zhang, C., Ji, Y., Zhao, Q., He, W., An, F., Jiang, L., and Guo, H. (2012). Activation of ethylene signaling is mediated by nuclear translocation of the cleaved EIN2 carboxyl terminus. Cell Res. *22*, 1613–1616.

Xu, J., and Chua, N.H. (2011). Processing bodies and plant development. Curr. Opin. Plant Biol. *14*, 88–93.

Zhang, X., Zhu, Y., Liu, X., Hong, X., Xu, Y., Zhu, P., Shen, Y., Wu, H., Ji, Y., Wen, X., et al. (2015). Suppression of endogenous gene silencing by bidirectional cytoplasmic RNA decay in Arabidopsis. Science *348*, 120–123.

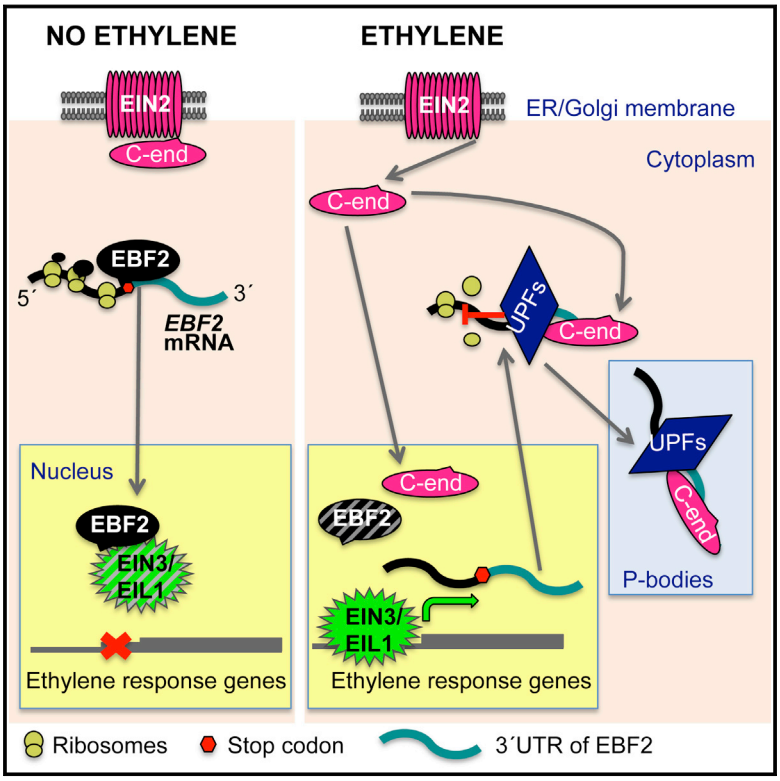
Zhao, W., Pollack, J.L., Blagev, D.P., Zaitlen, N., McManus, M.T., and Erle, D.J. (2014). Massively parallel functional annotation of 3′ untranslated regions. Nat. Biotechnol. *32*, 387–391.

Article

Cell

Gene-Specific Translation Regulation Mediated by the Hormone-Signaling Molecule EIN2

Graphical Abstract



Highlights

- Ribosome footprinting uncovers a role of translation in the ethylene response
- The *EBF2* 3′UTR is sufficient to confer translational control
- Regulation of *EBF2* translation is required for proper ethylene responses
- *EBF2* translation control depends on functional EIN2 and UPFs, but not EIN3/EIL1

Authors

Catharina Merchante, Javier Brumos, Jeonga Yun, ..., Steffen Heber, Anna N. Stepanova, Jose M. Alonso

Correspondence

atstepan@ncsu.edu (A.N.S.), jmalonso@ncsu.edu (J.M.A.)

In Brief

Ribosome footprinting unveils gene-specific regulation of translation by the hormone ethylene involving the 3′UTR of the transcript of a known negative regulator, as well as a key ethylene signaling protein and the components of the nonsense-mediated decay machinery.

Merchante et al., 2015, Cell *163*, 684–697
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.036

Cell

Article

Gene-Specific Translation Regulation Mediated by the Hormone-Signaling Molecule EIN2

Catharina Merchante,[1,4,6] Javier Brumos,[1,6] Jeonga Yun,[1] Qiwen Hu,[2] Kristina R. Spencer,[1] Paul Enríquez,[1] Brad M. Binder,[3] Steffen Heber,[2] Anna N. Stepanova,[1,5,*] and Jose M. Alonso[1,5,*]

[1]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27695, USA
[2]Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
[3]Department of Biochemistry, Cellular and Molecular Biology, University of Tennessee, Knoxville, Knoxville, TN 37996, USA
[4]IHSM-UMA-CSIC, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, 29071 Málaga, Spain
[5]Genetics Graduate Program, North Carolina State University, Raleigh, NC 27695, USA
[6]Co-first author
*Correspondence: atstepan@ncsu.edu (A.N.S.), jmalonso@ncsu.edu (J.M.A.)
http://dx.doi.org/10.1016/j.cell.2015.09.036

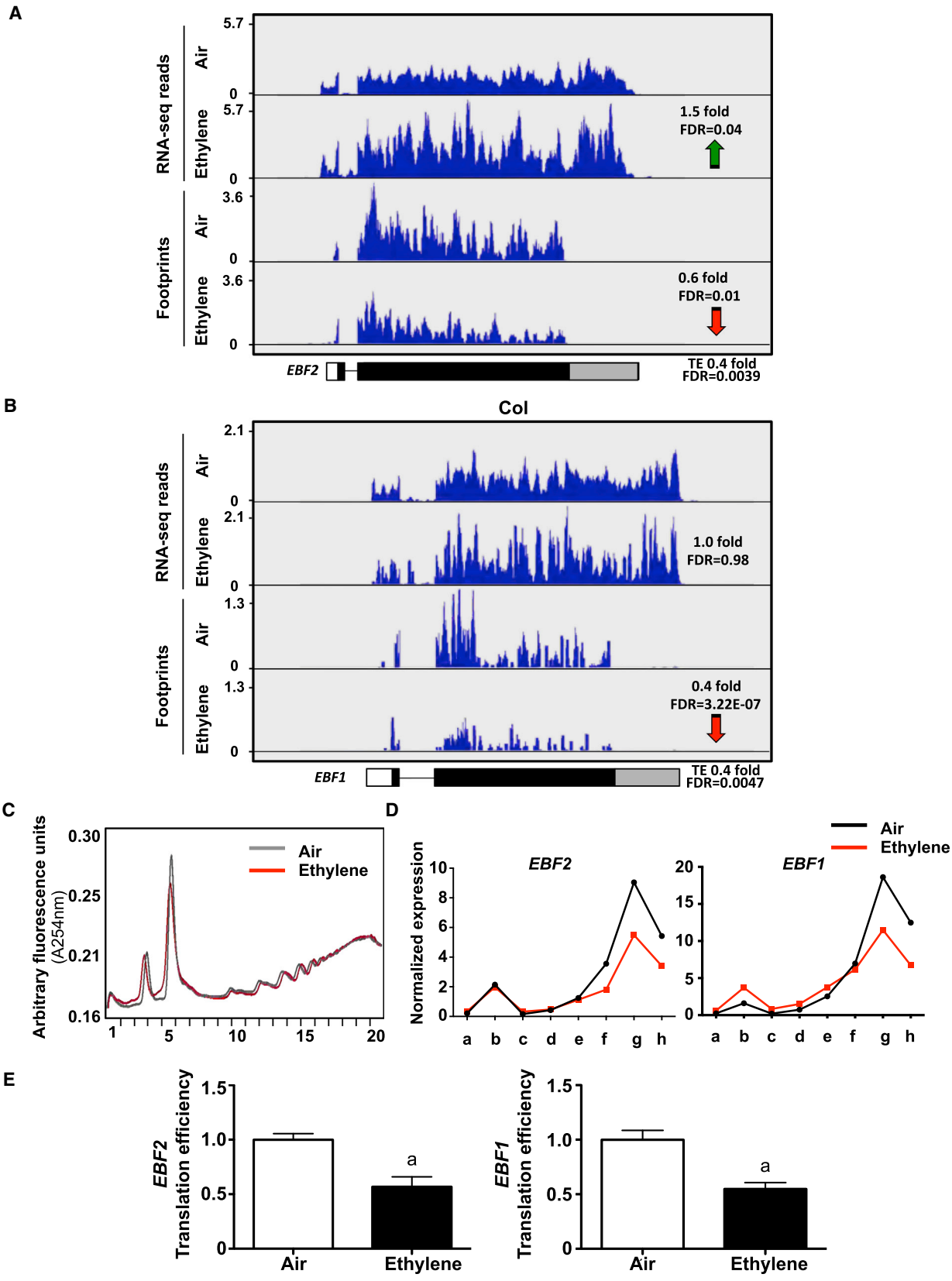
SUMMARY

The central role of translation in modulating gene activity has long been recognized, yet the systematic exploration of quantitative changes in translation at a genome-wide scale in response to a specific stimulus has only recently become technically feasible. Using the well-characterized signaling pathway of the phytohormone ethylene and plant-optimized genome-wide ribosome footprinting, we have uncovered a molecular mechanism linking this hormone's perception to the activation of a gene-specific translational control mechanism. Characterization of one of the targets of this translation regulatory machinery, the ethylene signaling component *EBF2*, indicates that the signaling molecule EIN2 and the nonsense-mediated decay proteins UPFs play a central role in this ethylene-induced translational response. Furthermore, the 3′UTR of *EBF2* is sufficient to confer translational regulation and required for the proper activation of ethylene responses. These findings represent a mechanistic paradigm of gene-specific regulation of translation in response to a key growth regulator.

INTRODUCTION

The plant hormone ethylene plays a central role in coordinating the multitude of molecular processes underlying developmental programs and environmental responses critical for plant survival (Abeles et al., 1992). The plant's response to ethylene is initiated by the binding of this hormone to its cognate receptors—in *Arabidopsis*, a small family of five proteins (ETR1, ETR2, ERS1, ERS2, and EIN4) with sequence similarity to the bacterial two-component histidine kinases (Bleecker et al., 1988; Hua and Meyerowitz, 1998). Although some specialization has been recognized for the receptors, they are all thought to function primarily by modulating the activity of the rapidly accelerated fibrosarcoma (RAF)-like kinase CTR1 (Clark et al., 1998). Inactivation of this kinase in the presence of ethylene results in a reduction in the phosphorylation levels of the endoplasmic-reticulum-localized transmembrane protein EIN2 and cleavage and translocation of the unphosphorylated C terminus of EIN2 (EIN2C) to the nucleus (Ju et al., 2012; Qiao et al., 2012). Downstream of EIN2, two different responses have been characterized. On the one hand, there is a rapid inhibition of growth that takes place within minutes of exposure to the hormone and does not involve the key transcriptional regulators EIN3 and EIL1 (Binder et al., 2004). On the other hand, there are many other, and possibly slower, changes induced by this hormone, including transcript level alterations in hundreds of genes that do require the function of these two transcriptional regulators (Binder et al., 2004; Chang et al., 2013). In contrast with the lack of information on the molecular mechanism behind the fast growth-inhibition response, all EIN3/EIL1-dependent responses are activated by the aforementioned translocation of the unphosphorylated EIN2C to the nucleus. Preventing this translocation stops the activation of EIN3/EIL1 (Qiao et al., 2012). The F-box proteins ETP1/ETP2 and EBF1/EBF2 control EIN2 and EIN3 protein abundance, respectively (Guo and Ecker, 2003; Potuschak et al., 2003; Qiao et al., 2009). Interestingly, *EBF2* itself is a direct transcriptional target of EIN3 (Konishi and Yanagisawa, 2008), suggesting the existence of a feedback regulatory loop that quickly dampens EIN3 activity shortly after activating this signaling cascade. The critical importance of the EIN3 regulation by the EBFs is further substantiated by the observation that EBF2 protein levels are also modulated by an unknown EIN2-dependent mechanism (He et al., 2011). Finally, a P-body-localized 5′-3′ exoribonuclease EIN5 (also known as XRN4) has also been implicated in the regulation of the *EBF2* activity (Olmedo et al., 2006; Potuschak et al., 2006; Souret et al., 2004; Weber et al., 2008).

Using a plant-optimized ribosome footprinting approach, we show that ethylene affects translation of several genes, among them the *EBF*s. The translational regulation of *EBF2* is mediated by its long 3′ UTR and requires the activity of the ethylene signaling components *EIN2* and *EIN5* and the nonsense-mediated decay proteins *UPF*s, but not that of the ethylene transcriptional master regulators *EIN3*/*EIL1*. EIN2C can interact with the 3′UTR of *EBF2* and localizes to P-bodies. These findings not only provide direct evidence for the translation regulation of



(legend on next page)

specific genes in response to this hormone but also the conceptual framework to decipher the molecular mechanism of a previously proposed branch of ethylene signaling.

RESULTS

Ribosome Footprinting Unveils a New Translation-Based Branch of the Ethylene Response

To probe the effects of ethylene on translation at the whole-genome level, we implemented the ribosome footprinting technology, Ribo-seq, which allows for capturing the ribosomal load of expressed genes in the genome at a single-codon resolution (Ingolia et al., 2009). Using Ribo-seq, we looked for ethylene-triggered changes in translation rates that could not be explained by changes in transcript levels.

Total mRNA and ribosome footprint analyses were carried out in parallel to identify changes in translation efficiency in response to ethylene (Figure S1A) (see the Supplemental Experimental Procedures). A 4 hr ethylene treatment was selected to capture robust early responses and to avoid secondary long-term effects of this hormone. The high quality of the Ribo-seq data (Ingolia et al., 2009) is evidenced by the abrupt appearance of a footprint signal 14–15 nt upstream of the start codon, a rapid decline in signal around 14–15 nt upstream of the stop codon, low density of footprints in the 5′ and 3′ UTR, and a strong 3 nt periodicity (Figures S1B–S1D), which represents the codon-long stepwise movement of the ribosome along the mRNA. None of these features were observed in the RNA sequencing (RNA-seq) libraries (Figures S1B–S1D).

Ethylene induced global mRNA level changes (Figure S1E and Table S1) that were followed by concomitant alterations in the levels of translation (Figures S1E and S1F and Table S1). However, in agreement with previous comparisons between protein and RNA levels (de Godoy et al., 2008; Ingolia et al., 2009), the correlation between the changes in transcript accumulation and translation levels was relatively poor, with an r^2 value of 0.22 (Figure S1F), suggesting the existence of a layer of regulation at the translational level. In fact, we identified several mRNAs affected by ethylene in their translational efficiency (Table S1). Importantly, two key ethylene signaling genes, *EBF1* and *EBF2*, were found in this list of translationally regulated genes (Table S1). *EBF1* and *EBF2* encode F-box proteins involved in the degradation of EIN3/EIL1 in the absence of ethylene. In prior studies, the EBF protein levels have been shown to decrease after ethylene treatment (Guo and Ecker, 2003; Potuschak et al., 2003), although the transcript levels of at least *EBF2* are known to increase in response to this hormone (Konishi and Yanagisawa, 2008). After 4 hr of exposure to ethylene, and coinciding with previous reports, we observed an ~1.5-fold increase in the *EBF2* mRNA, yet a surprising 2.8-fold decrease in its translation efficiency (TE) (Figure 1A and Table S1). Likewise, we observed a reduction in the TE of *EBF1* (Figure 1B) and several other genes (Figure S2 and Table S1). These ethylene effects on the translation of *EBF1* and *EBF2* were further supported by the reduction of the relative levels of these mRNAs in the heavy fractions of a polysome profile (Figures 1C and 1D). To further validate these findings, the ethylene effects on TE of six selected genes, including *EBF1*, *EBF2*, and a negative control, *RTE1*, were evaluated by calculating the ratio between the expression level of these genes in polysomal and total mRNA (Figures 1E and S2D). Although this approach is not as sensitive at detecting changes in the ribosomal load of an mRNA as are Ribo-seq or ribosome profiling, it can accurately quantify alterations in the ratio of the mRNA subpopulations that are actively engaged in translation versus those populations that are non-translating. The TE of *EBF1* and *EBF2* in ethylene decreased nearly to half of that in air (Figure 1E), confirming the results of Ribo-seq (Figures 1A and 1B) and polysome profiling (Figures 1C and 1D). Similarly, the TE of the other three selected genes was also repressed by ethylene, whereas no effect was detected for *RTE1*, a transcriptionally induced negative control (Figure S2D and Table S1). Together, these results suggest that the multitude of responses triggered by the hormone ethylene is the result of regulation of gene expression not only at the transcriptional level as shown previously (Chang et al., 2013) but also at the translational level. These changes in translation are likely due to shifts in the equilibrium of translated and non-translated populations of target mRNAs rather than quantitative alterations in the translation rates of individual transcripts. These findings also reveal that, as in the case of the transcriptional regulation, some of the components of the ethylene signal transduction pathway are themselves subject to ethylene-triggered translational regulation, raising the possibility of intricate feedback regulatory loops functioning in this signaling pathway.

The 3′UTR of *EBF2* Is Sufficient to Confer Ethylene-Mediated Regulation of Translation and Is Required for Proper Plant Responses to This Hormone

Since EBF2 is a key negative regulator of ethylene signaling (Guo and Ecker, 2003; Potuschak et al., 2003), we reasoned that the observed translational repression of this gene may have a significant physiological effect. Although translation regulatory

Figure 1. Translation of *EBF2* and *EBF1* Is Quickly Downregulated by Ethylene

(A and B) Normalized distribution of RNA-seq and Ribo-seq reads in air and in ethylene along the *EBF2* (A) and *EBF1* (B) genes.

5′ UTR, coding DNA sequence (CDS), 3′ UTR, and introns are marked as white, black, and gray boxes and a line, respectively. The fold change and the associated false discovery rate (FDR) for the ethylene effect on transcript and footprint levels, as well as the fold change in the footprint levels given the levels of mRNA (TE) and the corresponding FDR, are shown.

(C) 10%–50% sucrose gradient absorbance (A_{254}) profiles of ribosome complexes obtained from *Arabidopsis* seedlings grown in air and/or 4 hr ethylene.

(D) Polysomal distribution of *EBF* transcripts in air and 4 hr ethylene. a–h correspond to fractions 4+5 through 18+19 shown in (C) pooled in pairs. *EBF* mRNA levels were normalized against *At4g34270*. *EBF* expression in each polysomal fraction was calculated as the percentage of its expression in total RNA.

(E) TE of the *EBF2* and *EBF1* mRNAs, calculated as their relative expression in polysomal/total RNA fractions, in seedlings grown in air or treated with 10 ppm of ethylene for the last 4 hr of the experiment. Expression levels of *EBFs* were normalized against *At4g34270*. (a) indicates a significant difference of the ethylene effect on the *EBF* TE (t test, $p < 0.05$). Bars represent means ± SEM for three biological replicates.

3-day-old etiolated seedlings were used in all of the experiments.

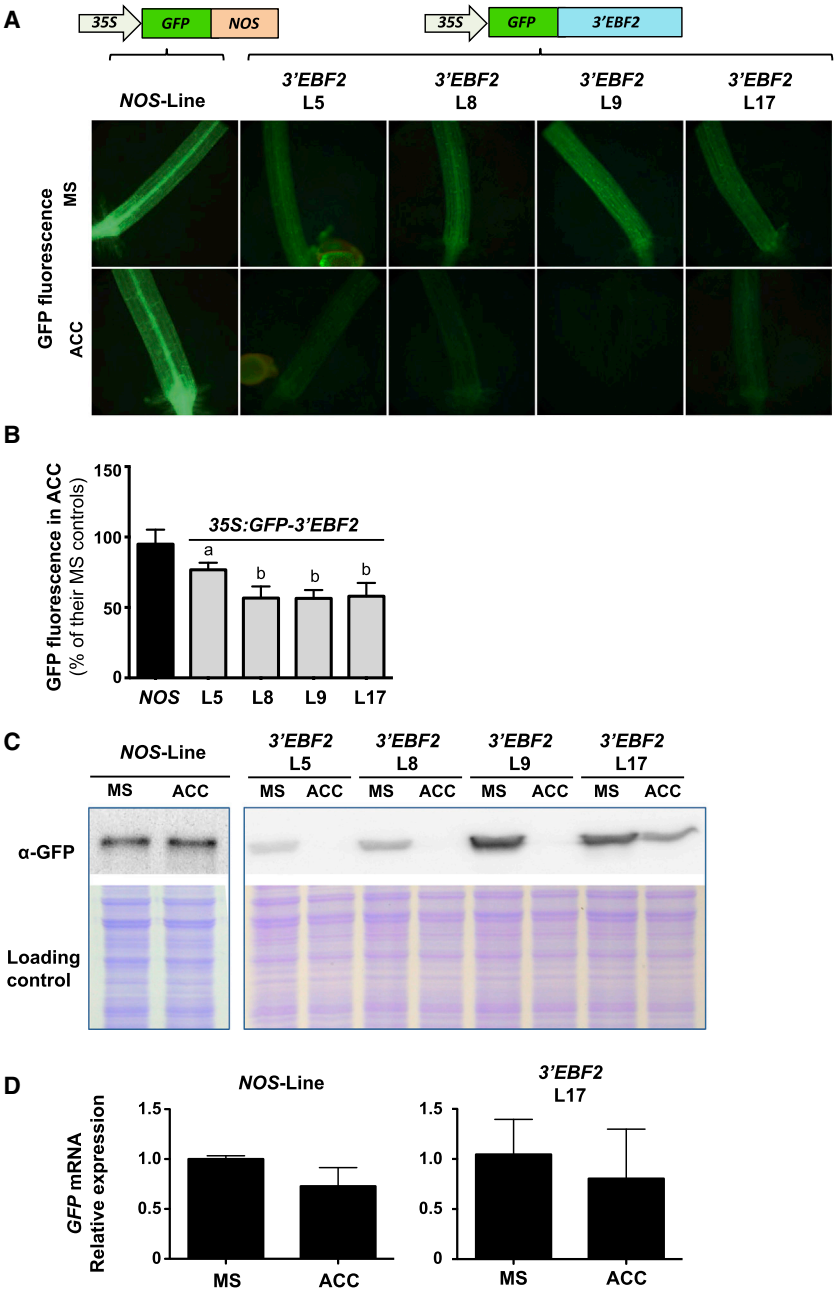


Figure 2. The *3′EBF2* Is Sufficient to Confer Ethylene-Mediated Regulation of Translation

(A and B) (A) Hypocotyl fluorescence and (B) its quantification (n = 15) in 3-day-old etiolated seedlings grown in the presence (ACC) or absence (MS) of the ethylene precursor ACC and harboring either the *35S:GFP-NOS* or the *35S:GFP-3′EBF2* constructs as depicted on top of the photos. GFP fluorescence is expressed as the % of fluorescence in ACC compared to that in MS controls. Bars represent means ± SD. a and b indicate a significant effect of the ethylene treatment on the levels of fluorescence (t test, $p < 0.005$ and $p < 0.0001$, respectively).

(C) Anti-GFP western blot of total protein extracts from the transgenic lines shown in (A).

(D) Relative expression of *GFP* mRNA from two selected lines from (A). Bars represent means ± SEM for three biological replicates. Expression levels of the *EBF2* transgenes were normalized against *At5g44200*.

3-day-old etiolated seedlings were used in all of the experiments.

elements can be located in both 5′ and 3′ UTRs (Szostak and Gebauer, 2013), we decided to investigate the potential regulatory role of the atypically large 590-bp-long 3′ UTR of *EBF2* (*3′EBF2*) first, as it has previously been implicated in modulating the activity of this gene and prior efforts to determine the mechanism of such regulation via changes in *EBF2* mRNA stability were not conclusive (Potuschak et al., 2006). *GFP* reporter was fused to either *3′EBF2* or the *NOS* terminator and placed under the control of the constitutive *35S* promoter (Figure 2A). The effect of ethylene on the GFP fluorescence of stably transformed wild-type plants was examined under the standard ethylene triple response assay conditions. While the transgenic lines with the *NOS* terminator alone were equally fluorescent in control media and in media supplemented with the ethylene precursor ACC, the *3′EBF2* lines showed a strong reduction in the levels of fluorescence in ACC (Figures 2A and 2B). Western blot with an anti-GFP antibody confirmed that the observed decrease in fluorescence in the latter was due to a reduction in the amount of GFP protein (Figure 2C), whereas the *NOS* terminator line showed equal amounts of GFP protein in the presence and absence of ACC. As expected, we observed a range of GFP protein levels in different transgenic lines, and in general, lower levels were found in the *3′EBF2* lines even in the absence of exogenous ethylene, observations that likely resulted from positional effects and endogenous ethylene. To further demonstrate that this ethylene effect in the *3′EBF2* lines was due to changes at the level of protein translation rather than transcription or mRNA stability, the *GFP* mRNA was quantified by qRT-PCR. The differences in the GFP protein accumulation could not be explained by an ethylene-mediated effect on the mRNA levels (Figure 2D), which is consistent with previous reports that did not detect an effect of ethylene on the mRNA stability of *EBF*s (Potuschak et al., 2006).

To better understand the role of *3′EBF2* in mediating the observed ethylene effect on translation, we took a complementary approach utilizing previously generated transgenic lines (Konishi and Yanagisawa, 2008). In these lines, the *ebf2* mutant is complemented with either a native genomic construct of *EBF2* or a similar construct in which *3′EBF2* was replaced by the *NOS* terminator (Figure 3A). Transgenic lines complemented with the native genomic construct showed a clear reduction in the TE of

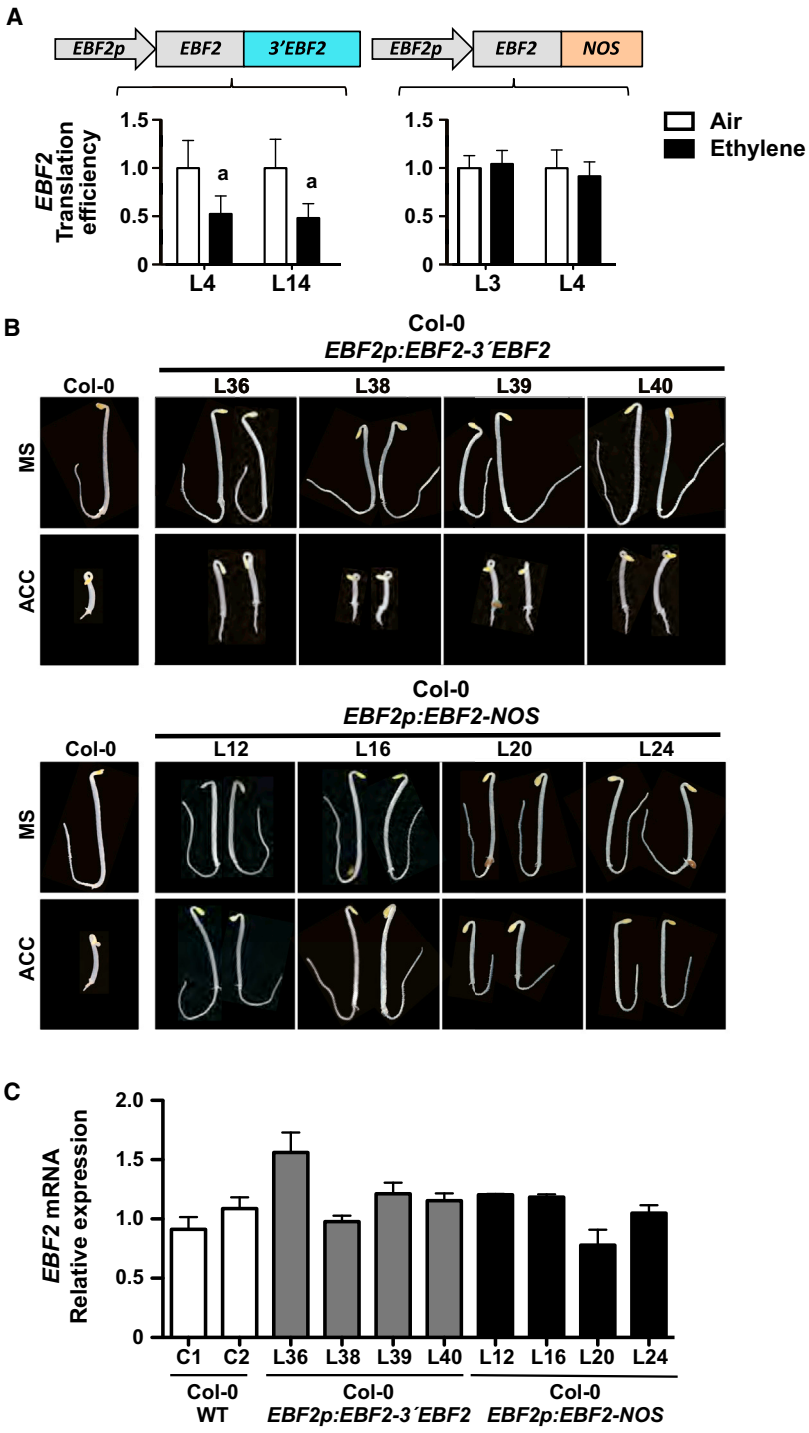


Figure 3. The *3′EBF2* Is Required for the Proper Translation of *EBF2* and Plant Response to Ethylene

(A) TE of the *EBF2* mRNA, calculated as the relative expression in polysomal/total RNA fractions, in *ebf2* seedlings complemented with *EBF2p:EBF2-3′EBF2* and *EBF2p:EBF2-NOS* constructs grown in air or treated with 10 ppm of ethylene for the last 4 hr of the experiment. (a) indicates a significant difference of the ethylene effect in the different genotypes (ANOVA, $p < 0.01$). Bars represent means ± SEM for three biological replicates. Expression levels of the *EBF2* transgenes were normalized against *At4g34270*.

(B) Representative images of 3-day-old etiolated seedlings of the indicated genotypes grown in control media (MS) or in the presence of 10 μM ACC (ACC).

(C) *EBF2* mRNA expression levels in 3-day-old etiolated seedlings of the same genotypes as shown in (B) normalized against *At5g44200* and expressed relative to the average of Col-0 plants.

Error bars represent means ± SEM for three technical replicates.

EBF2 (Figure 3A), equivalent to that of the native *EBF2* in wild-type plants (Figure 1). However, ethylene had no effect on the TE of *EBF2* in the *NOS* terminator lines (Figure 3A). Taken together, these results demonstrate that *3′EBF2* is sufficient to confer ethylene-mediated translational regulation.

We re-examined the ethylene response of the aforementioned *ebf2* lines expressing either the native *EBF2* genomic construct or the corresponding control with the *3′EBF2* replaced by the *NOS* terminator (*EBF2-NOS* lines). As previously reported (Konishi and Yanagisawa, 2008), the lines with the native *3′EBF2* showed normal ethylene response in the classical triple response assay, whereas the *NOS* terminator lines showed strong ethylene insensitivity (Figure S3A). Although previously these phenotypes were attributed to the slightly elevated levels of *EBF2* mRNA in the *NOS* terminator lines (Konishi and Yanagisawa, 2008), our results suggested that the ethylene insensitivity of these lines could also be due to the loss of the 3′ UTR-mediated translational repression of *EBF2* by ethylene (Figure 3A). To distinguish between these two possibilities, we generated additional transgenic lines in a wild-type genetic background using either the native *EBF2* or the *EBF2-NOS* terminator constructs. As shown in Figures 3B and 3C, no correlation between the ethylene phenotype and the levels of *EBF2* mRNA was observed (Figures 3B and 3C). The biological significance of the regulatory role of *3′EBF2* was further supported by the observation that plants expressing *GFP-3′EBF2* under the strong *35S* promoter displayed mild ethylene insensitivity (Figure S3B). These results suggest that the presence of high levels of *3′EBF2* can interfere with the molecular machinery responsible for the translational repression of the endogenous *EBF2* mRNA. Taken together, the findings described above strongly support the idea that the translational regulation conferred by *3′EBF2* is critical for the proper function of the ethylene signaling pathway and the plant response to this hormone.

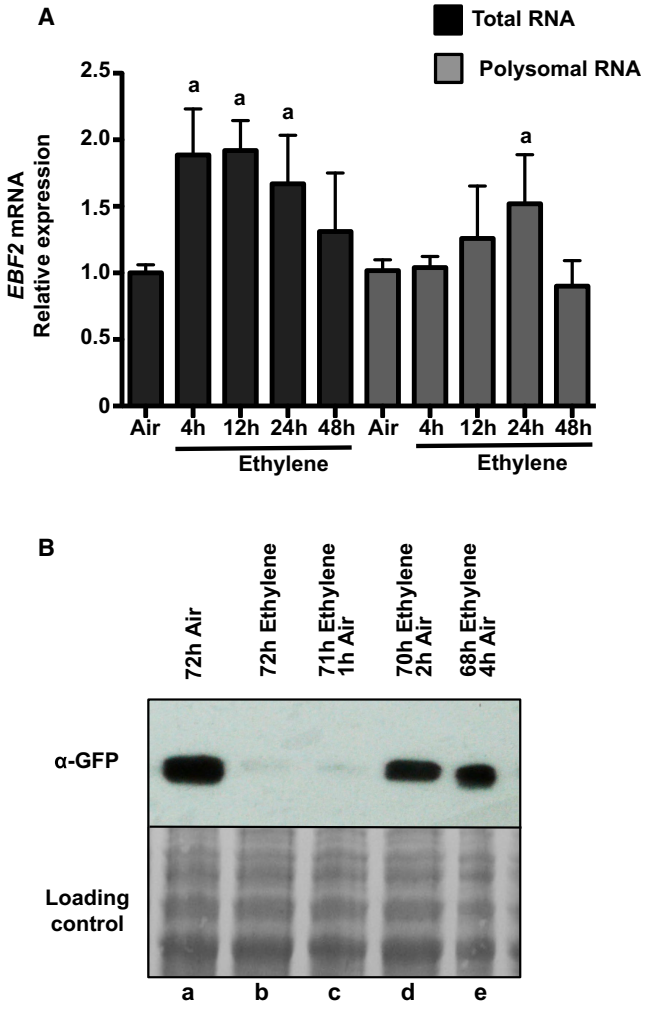


Figure 4. *EBF2* Displays Complex Transcriptional and Translational Dynamics in Response to Ethylene

(A) Relative expression levels of *EBF2* mRNA in total and polysomal RNA fractions from 3-day-old etiolated Col-0 seedlings during a time-course treatment using 10 ppm of ethylene. (a) indicates significant difference between that time point and the corresponding "Air" control (t test, $p < 0.05$). Bars represent means ± SEM for three biological replicates. Expression levels of *EBF2* were normalized against *At4g34270*.

(B) Anti-GFP western blot in *35S:GFP-3′EBF2* of total protein extracts from 3-day-old etiolated seedlings during a time-course ethylene withdrawal experiment.

The Dynamics of Transcriptional and Translational Regulation of *EBF2* Shed New Light on Molecular Mechanisms of the Ethylene Response Kinetics

Our results regarding the opposite effects of a short ethylene exposure on *EBF2* expression at the transcriptional and translational level, together with the known role of *EBF2* in the control of *EIN3* activity, suggest existence of a regulatory mechanism involved in the dynamic aspects of the ethylene response. To investigate the role of transcriptional and translational regulation of *EBF2* in the observed kinetics of the ethylene response, we examined by qRT-PCR the levels of *EBF2* mRNA in both total and polysomal RNA fractions at different times after initiating the ethylene treatment. In agreement with previous reports (Chang et al., 2013), the levels of *EBF2* mRNA quickly increased, reaching the highest expression in the total RNA sample 4 hr after the beginning of the ethylene treatment, staying high at the 12 hr time point and slowly decreasing thereafter (Figure 4A). In contrast, polysomal *EBF2* mRNA remained low for the first 4 hr (Figure 4A), despite the high levels of total *EBF2* mRNA. Lack of efficient translation of *EBF2* (a negative regulator of ethylene responses that targets EIN3 for degradation) thus allows for a full-scale ethylene response in early stages of exposure of plants to ethylene. Interestingly, this period of low EBF2 accumulation coincides with the previously reported maximum in EIN3 activity (Chang et al., 2013). Only after a prolonged ethylene exposure (12 hr to 24 hr) did we observe an increase in *EBF2* mRNA accumulation in the polysomal fraction (Figure 4A), which correlates with the previously described decrease in EIN3 activity (Chang et al., 2013) and coincides with a parallel decline in the total mRNA levels of *EBF2* (Figure 4A) (Chang et al., 2013). Thus, the attenuation of the ethylene response under continuous exposure to this hormone is preceded by an increase in the mRNA levels of *EBF2* in the polysomal fraction (Figure 4A), suggesting that the dynamic balance between transcriptional and translational activity of *EBF2* plays a critical role in diminishing the ethylene response upon prolonged exposure to the hormone.

To examine the reversibility of the ethylene effect on translation, we performed a time-course recovery experiment using the *p35S:GFP-3′EBF2* lines (Figure 4B) that can monitor the ethylene effect specifically on translation—i.e., in the absence of transcriptional regulation. We compared the accumulation of the GFP protein in seedlings grown in air, exposed to ethylene for the entire duration of the experiment (72 hr), or exposed to ethylene for 71, 70, or 68 hr and then allowed to recover in the absence of the hormone for the last 1, 2, or 4 hr of the total 72-hr-long experiment, respectively (Figure 4B). As shown in Figure 4B, in spite of the attenuation process described above, ethylene was able to nearly completely suppress the translation of the *3′EBF2*-containing mRNA expressed under a strong constitutive promoter even after 72 hr of constant exposure to the hormone. Importantly, the protein levels of GFP rapidly increased after ethylene was removed, reaching maximum levels just 2 hr after the withdrawal of ethylene (Figure 4B). These results support the idea that the translation regulation conferred by *3′EBF2* plays a role in re-establishing homeostasis upon removal of ethylene. Importantly, the analysis of the *ebf2* mutant has previously implicated this gene in the resumption of growth after ethylene withdrawal (Binder et al., 2007), further supporting the physiological significance of the observed translation dynamics of this gene.

The Ethylene-Triggered Regulation of Translation of *EBF2* mRNA Is *EIN2* Dependent but *EIN3/EIL1* Independent

To determine which canonical components of this hormone signaling pathway are required to mediate the translational regulation of *EBF2* mRNA, we examined the expression of the *35S:GFP-3′EBF2* construct in the strong ethylene signaling mutants *ein2-5* and *ein3-1 eil1-1* (Figure 5). The GFP fluorescence of

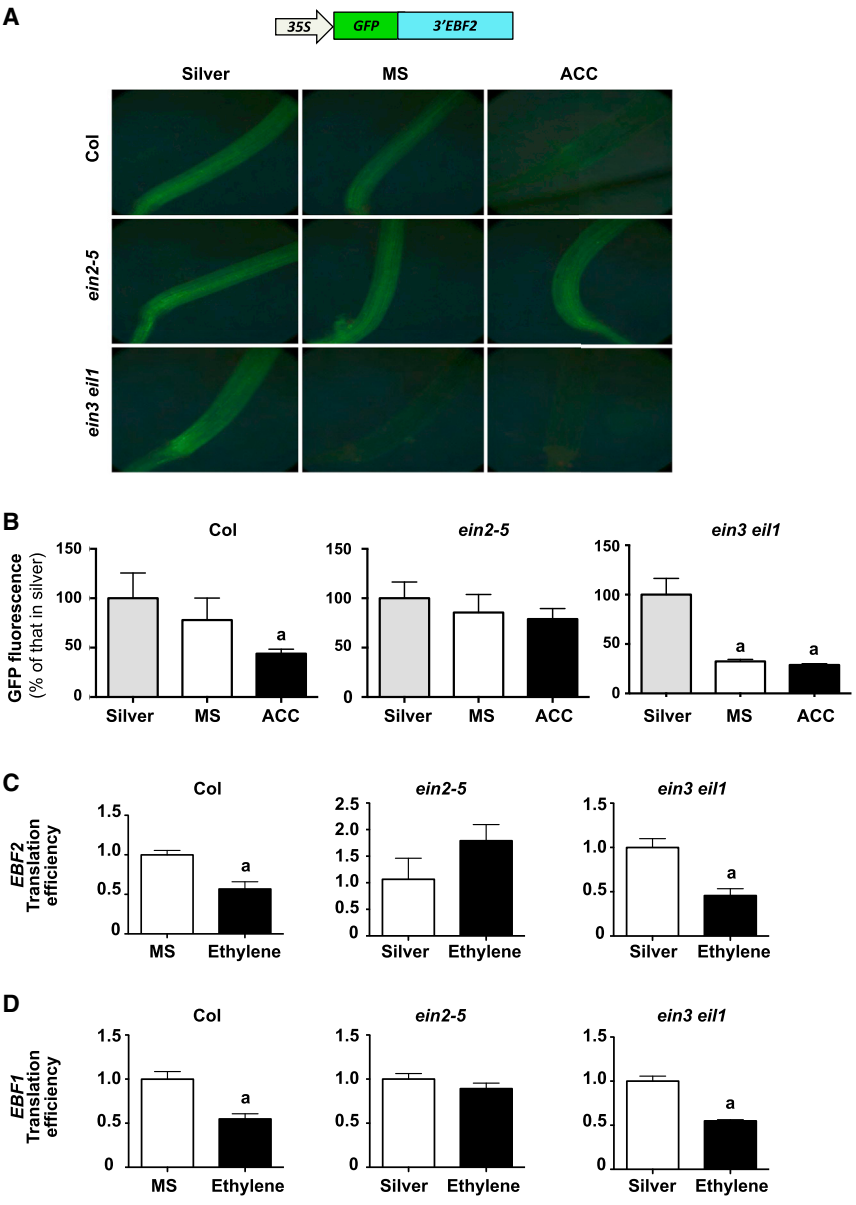


Figure 5. The Ethylene-Dependent Regulation of Translation of *EBF2* and *EBF1* Is *EIN2* Dependent but *EIN3/EIL1* Independent

(A–D) (A) Hypocotyl fluorescence, (B) its quantification, and (C and D) TE of the endogenous *EBF2* (C) and *EBF1* (D) mRNA in 3-day-old etiolated Col-0, *ein2-5*, and *ein3-1 eil1-1* seedlings harboring *35S:GFP-3′EBF2* and grown in 5 mg/l silver, air, or in 10 μM ACC.

(B and C) GFP fluorescence (B) was quantified across multiple seedlings (n = 7) and expressed as the percentage of fluorescence in ACC compared to that in silver controls. Error bars in (B) represent means ± SD. (a) indicates a significant effect of the ethylene treatment on the levels of fluorescence (one-way ANOVA, $p < 0.0001$). TE was calculated as the relative expression in polysomal/total RNA fractions. Col graphs are the same as those shown in Figure 1, plotted here again to facilitate their comparison with the mutants. (a) indicates a significant difference of the ethylene effect on the *EBF2* mRNA TE (t test, $p < 0.05$). Error bars in (C) and (D) represent means ± SEM for three technical replicates. Expression levels of the *EBF* transgenes were normalized against *At4g34270*.

wild-type, *ein2*, and *ein3 eil1* seedlings homozygous for the transgene was examined (Figures 5A and 5B). A silver-treated control was included to mitigate the effect of elevated endogenous levels of ethylene in *ein2* and *ein3* mutants (Guzmán and Ecker, 1990; Vandenbussche et al., 2012). While ethylene had a dramatic effect on the levels of GFP fluorescence in the wild-type plants, we did not observe any changes in the GFP intensity in the *ein2* plants treated either with silver or with the ethylene precursor ACC (Figures 5A and 5B). Surprisingly, the levels of fluorescence in the *ein3 eil1* double mutant were clearly affected by ethylene (Figures 5A and 5B). GFP fluorescence in this mutant was high in silver (where the effect of endogenous ethylene was suppressed) but dramatically decreased in plants grown in ACC-supplemented or un-supplemented media (where the high levels of endogenous ethylene were sufficient to trigger a full response). These results indicate that, while the function of *EIN2* is required for the *3′EBF2*-mediated ethylene-triggered regulation of translation, *EIN3/EIL1* are not.

Next, we investigated the effects of the *ein2* and *ein3 eil1* mutants on the TE of the endogenous *EBF2* and *EBF1* in plants grown in silver- or ACC-supplemented media (Figures 5C and 5D). In full agreement with the results obtained for the GFP fluorescence of the *35S:GFP-3′EBF2* reporter construct, the TE of *EBF*s did not significantly change in the presence of silver or ACC in the *ein2* mutant, but a robust reduction in TE typically observed in wild-type plants in response to ethylene was also seen in the *ein3 eil1* mutant (Figures 5C and 5D).

To determine the molecular mechanism by which *EIN2* mediates translational repression of the *EBF2* mRNA, both a miRNA-based and a protein/RNA interaction-based mechanisms were considered. A negative outcome of prior efforts to elucidate the possible regulation of *EBF2* by miRNA (Souret et al., 2004), together with the lack of known or predicted miRNA in the *Arabidopsis* genome likely to target *EBF2* (Alves et al., 2009), made us disfavor the miRNA possibility. Nevertheless, we decided to examine the ethylene response of the strong small RNA biogenesis mutant *dcl2-1 dcl3-1 dcl4-2* (Henderson et al., 2006). Consistent with the idea that small RNAs are not involved in the regulation of translation of *EBF2*, the mutant displayed wild-type level of ethylene sensitivity in the standard triple

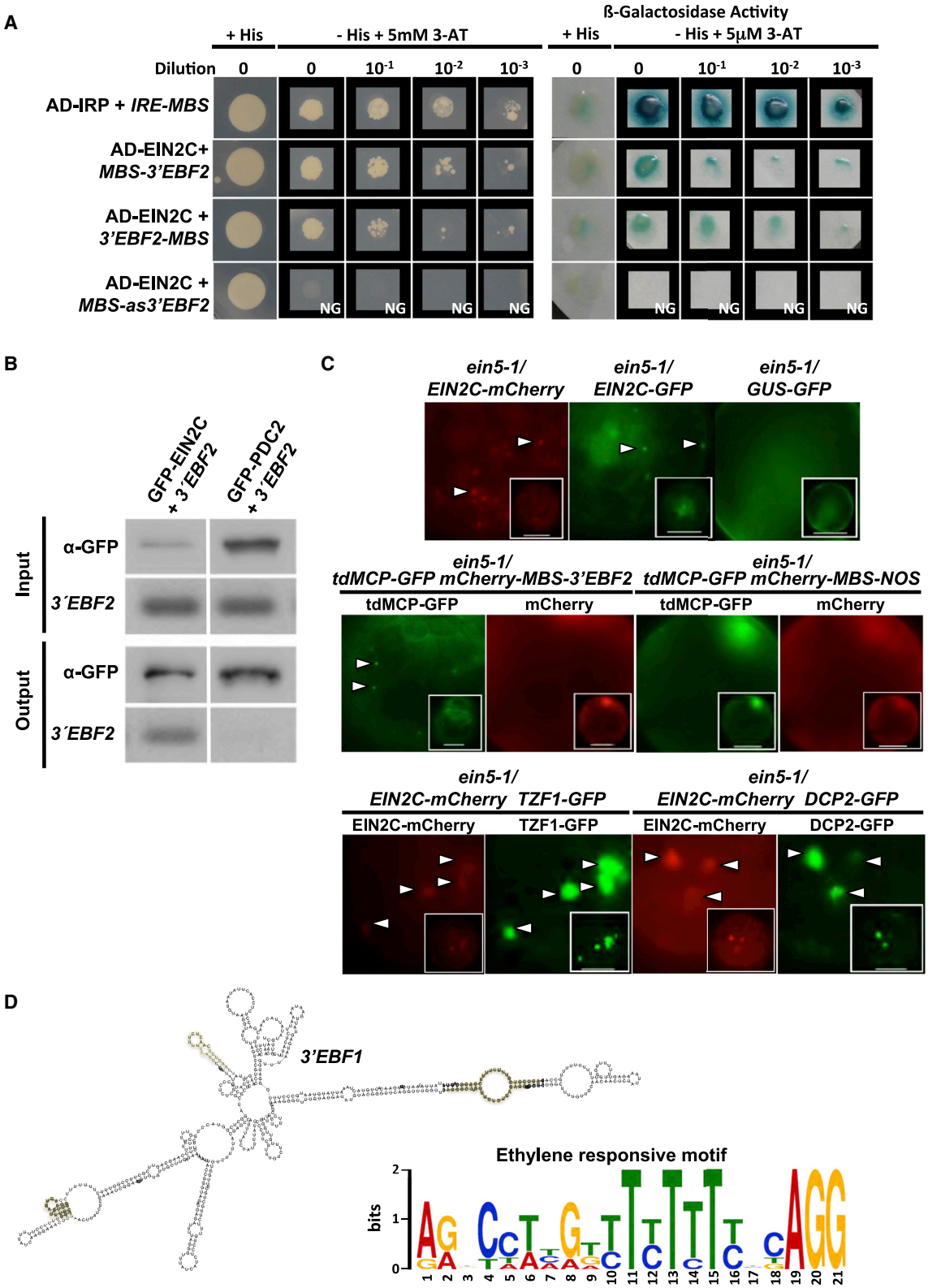


Figure 6. The EIN2C Interacts with the *3′EBF2* mRNA and Localizes to P-Bodies

(A) Yeast three-hybrid assay of the interaction between *3′EBF2* and EIN2C. Activity of the reporter genes for interaction between the RNA bait and the protein prey is shown (HIS3 activity [growth] on His- media, left, and β-galactosidase [blue color] in X-gal, right). All yeast strains employed harbor the DNA binding domain of

(legend continued on next page)

response assay (Figure S3C). These results are in agreement with the previous observation that *hen1*, *rdr2*, *dcl2*, *dcl3*, *sde1*, and *dcl4* mutants are not impaired in their response to ethylene (Potuschak et al., 2006).

Since *EIN2* is the most downstream known signaling component required for the translational regulation of *EBF2*, we decided to examine the interaction between the signal transducer—i.e., EIN2C (Alonso et al., 1999)—and the *3′EBF2* mRNA. Using the yeast three-hybrid system (SenGupta et al., 1996), we were able to detect interaction of EIN2C with two different RNA hybrids of *3′EBF2*, but not with the antisense version of this 3′ UTR (Figure 6A). Next, we investigated if EIN2C from plant extracts could bind its target *EBF2* mRNA in vitro. Using *Nicotiana benthamiana*, we transiently expressed *35S:GFP-EIN2C* or the negative control *35S:GFP-PDC2* and measured the capacity of the tagged proteins to bind to the in-vitro-transcribed *3′EBF2* RNA in an RNA-immunoprecipitation assay (Figure 6B). While we could detect the RNA for *3′EBF2* in the samples with EIN2C, we were not able to detect its presence in the PDC2 control samples. These results suggest that EIN2C could bind to mRNAs in the cytoplasm and regulate their translation. An obvious implication of this mechanistic model is that EIN2C should be localized not only in the nucleus, as previously reported (Qiao et al., 2012), but also in the cytosol. To test this possibility, we reexamined the subcellular localization of transiently expressed GFP- or mCherry-tagged EIN2C in *Arabidopsis* protoplasts and/or tobacco leaves. As reported previously, EIN2C was nuclear localized (Figures S4A and S4B). We reasoned that, perhaps, under standard conditions, only a small fraction of EIN2C and/or only transiently is localized in the cytosol. Two different approaches were used to enhance the activity of EIN2C in the cytosol. First, we examined the subcellular localization of (1) GFP-tagged EIN2C in tobacco leaves co-transfected with a construct expressing *mCherry-3′EBF2* under the strong *35S* promoter (Figure S4B) and of (2) mCherry-tagged EIN2C in *Arabidopsis* protoplasts co-transfected with a construct expressing *CFP-MBS-3′EBF2* under the strong *35S* promoter (Figure S4C). While, in *Arabidopsis* protoplasts, we were not able to consistently detect a significant alteration in the EIN2C subcellular distribution, with EIN2C detected mainly in the nucleus (Figure S4C), in tobacco leaves, EIN2C was consistently localized both in the nucleus and in the cytoplasm where it formed distinct fluorescent foci (Figure S4B). Next, we examined the subcellular localization of mCherry- or GFP-tagged EIN2C in protoplasts obtained from the *Arabidopsis ein5-1* mutant known to accumulate high levels of *3′EBF2* (Potuschak et al., 2006). As shown in Figure 6C, the subcellular distribution of tagged EIN2C in *ein5* dramatically shifted from nuclear to dual nuclear/cytoplasmic localization. As in tobacco, EIN2C in *ein5* was not uniformly distributed in the cytosol but rather formed punctate aggregates. In contrast, localization of the GFP fusion protein expressed from the control *35S:GUS-GFP* construct was not affected by the *ein5* mutation (Figures 6C and S4D). The EIN2C aggregates were found to correspond to P-bodies by co-localization experiments between EIN2C-mCherry and the P-body markers TZF1-GFP and DCP2-GFP (Goeres et al., 2007; Pomeranz et al., 2010) (Figure 6C). Furthermore, *3′EBF2* localized to similar cytoplasmic granules both in *ein5-1* protoplasts (Figure 6C) and wild-type tobacco leaves (Figure S4E). These results not only support the idea that EIN2C is part of an RNA-protein complex localized to the P-bodies but also suggest that the ethylene defects of *ein5* could be the consequence of an overload of the translation regulation machinery similar to what is observed in plants overexpressing *GFP-3′EBF2* under the strong *35S* promoter (Figure S5). Alternatively, or perhaps in addition to this effect, the ethylene insensitivity of *ein5* could also be caused by the disruption of the normal trafficking of EIN2C to the nucleus. Consistent with the idea that, in *ein5* the translation regulatory machinery involved in the control of *EBF2* mRNA translation is overloaded and, therefore, defective, we observed that, in the *ein5* mutant, the effect of ethylene on the GFP fluorescence level of *35S:GFP-3′EBF2* and on the TE of the endogenous *EBF2* mRNA was significantly reduced compared with the responses observed in the corresponding wild-type controls (Figure S5).

Having established that the ethylene responsive element mediating the EIN2-dependent translation regulation is located in the 3′UTRs of *EBF2* and *EBF1* mRNAs, we searched for a conserved sequence motif. Using the MEME motif finder (Bailey and Elkan, 1994), a conserved motif present multiple times in these two genes was identified (Figure 6D). Importantly, using the AME package (McLeay and Bailey, 2010), this motif was shown to be significantly enriched (p value = 2.63e−3) among the 3′ UTRs of the genes translationally regulated by ethylene (Table S1).

Translation Regulation and Ethylene Responses Are Disrupted in the *upf2* Mutants

In a parallel approach to identify additional genes involved in the ethylene response, five mutants were found and ordered in three complementation groups (Figures 7A, and S6A, and S6B). Map-based cloning and identification of additional insertional alleles showed that the causal mutations resided in the three core components of the nonsense-mediated RNA decay machinery, *UPF1*, *UPF2*, and *UPF3* (Figure S6B and Table S3). Based on

LexA fused to the bacteriophage MS2 coat protein (MCP). The additional constructs specific for each strain are the positive control prey protein (AD-IRP), the positive control RNA bait (*IRE-MBS*), the GAL4-activation domain fused to EIN2C (AD-EIN2C), and three different *3′EBF2* RNA baits with the *MS2 binding site* (*MBS*) fused in the sense orientation to the 5′ (*MBS-3′EBF2*) and 3′ (*3′EBF2-MBS*) ends or in the antisense orientation (*MBS-as3′EBF2*). NG indicates no growth.

(B) RNA immunoprecipitation of GFP-EIN2C and the control protein GFP-PDC2 purified from transfected tobacco leaves and incubated with the in-vitro-transcribed *3′EBF2* mRNA. The protein and RNA levels in the input and those retained in the anti-GFP column are shown.

(C) Representative images of *ein5-1* mesophyll protoplasts transfected with the indicated constructs. TZF1 and DCP2 were used as P-body markers. tdMCP is the tandem version of MCP. *MBS* corresponds to 24 copies of the MCP binding sequence. The white scale bar represents 25 μm.

(D) Predicted ethylene-responsive translation *cis*-regulatory element. MEME motif finder identified a consensus sequence present in the 3′ UTRs of *EBF1* and *EBF2* mRNAs and significantly enriched in the 3′UTRs of genes regulated at the translational level by ethylene. The sequence of the ethylene responsive motif (inner panel) and the secondary structure of the 3′UTR of *EBF1* (with the motif-matching sequences highlighted in yellow) are shown.

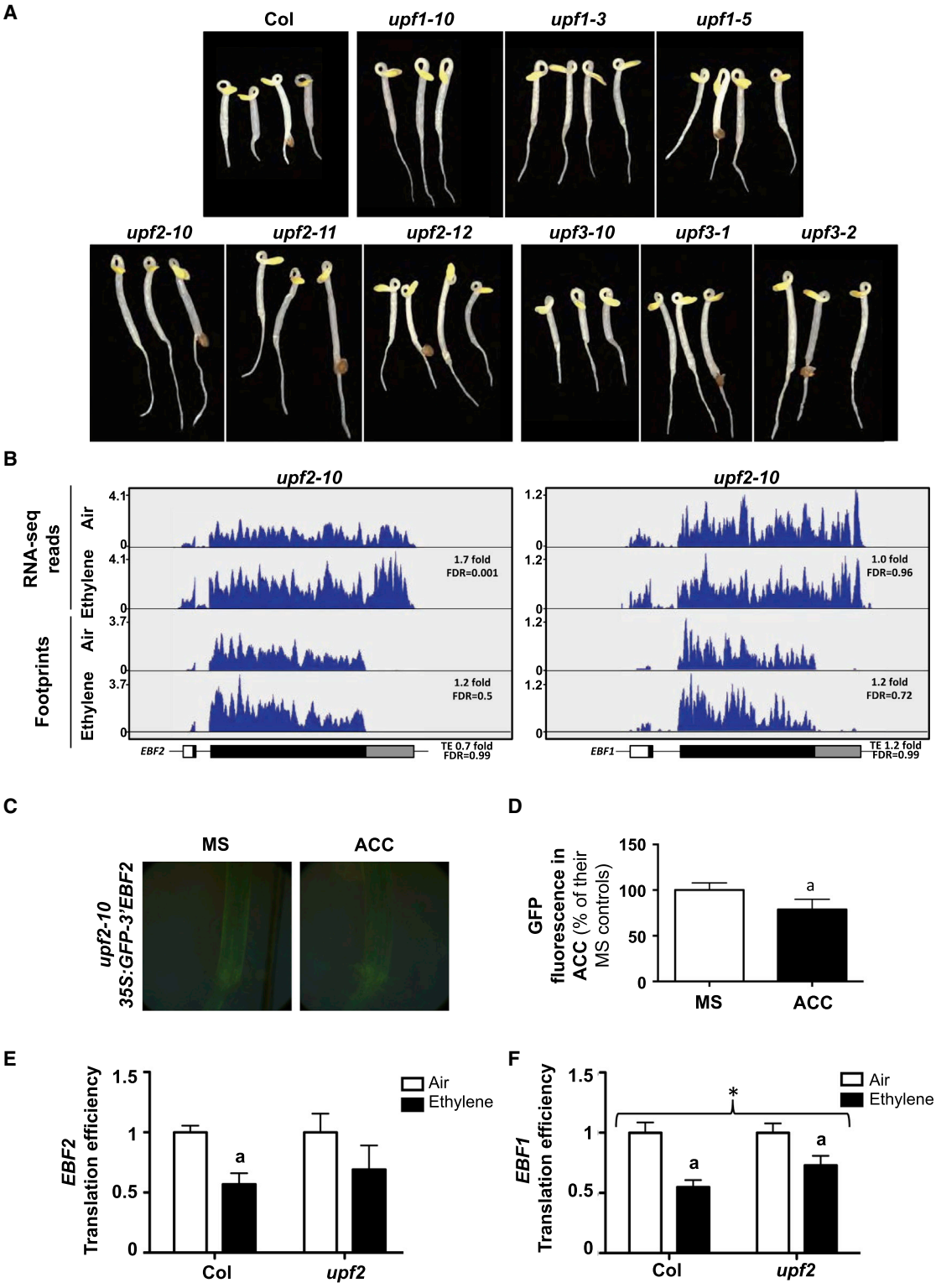


Figure 7. Ethylene-Triggered Root Growth Inhibition and Translational Regulation of *EBF*s Are Disrupted in the *upf2* Mutant

(A) Representative images of 3-day-old etiolated seedlings of the indicated genotypes grown in the presence of 10 μM ACC.

(B) Normalized distribution of RNA-seq and Ribo-seq reads in air and in ethylene along the *EBF2* (left) and *EBF1* (right) genes in *upf2-10*. 5′ UTR, CDS, 3′ UTR, and intron are marked as white, black, and gray boxes and a line, respectively. The fold change and the associated FDR for the ethylene effect on transcript and footprint levels, as well as the fold change in the footprint levels given the levels of mRNA (TE) and the corresponding FDR, are shown.

(legend continued on next page)

our findings on the role of ethylene in the regulation of translation of *EBF2* mRNA through its atypically long 3′ UTR, the lack of ethylene effects on the *EBF* mRNA stability (Potuschak et al., 2006), and the known role of the UPFs in inhibiting translation and targeting to the P-bodies of mRNAs with long 3′ UTRs, we decided to investigate the function of the UPFs in the translational regulation of *EBF*s. Ribo-seq experiments in the hypomorphic allele of *UPF2*, *upf2-10*, clearly show that the translational regulation by ethylene of *EBF2*, *EBF1* (Figure 7B and Table S2), and several other ethylene-regulated mRNAs (Figure S2) was dramatically attenuated. Similarly, the analysis of GFP fluorescence from the *35S:GFP-3′EBF2* construct (Figures 7C and 7D) and the quantification of TE of the endogenous *EBF* mRNA in response to ethylene (Figures 7E and 7F) also show that the *upf2-10* mutation attenuates the ethylene-induced translational regulation of the *EBF* mRNA. These findings, together with our results that translational regulation mediated by EIN2 does not require functional EIN3/EIL1, suggest that the function of UPFs needed for proper ethylene response is also required upstream of EIN3/EIL1. Importantly, we observed that, as in the case of EIN2C, the subcellular localization of UPF1 is also altered in the *ein5* background (Figure S4F), changing from a uniform nuclear/cytosolic to markedly punctuated foci. This subcellular localization of UPF1 partially overlaps with that of EIN2C (Figure S4G), suggesting P-body localization. Finally, we analyzed the kinetics of the ethylene response and recovery of both *upf2* and the mild ethylene insensitive transgenic plants expressing *35S:GFP-3′EBF2*. This analysis shows that both the *upf2* mutant and the *35S:GFP-3′EBF2* transgenic lines show similar defects during the recovery process after ethylene exposure (Figures S3D and S6C).

DISCUSSION

The response of plants to the hormone ethylene has been extensively studied, and a linear signaling pathway responsible for triggering the multitude of responses to this hormone has been identified. Importantly, all known gene-expression changes triggered by this hormone require the entirety of the pathway, including *EIN2* and the master transcriptional regulators *EIN3* and *EIL1* (Chang et al., 2013; Olmedo et al., 2006). Other facets of the ethylene response, however, have been shown to require the activity of *EIN2*, but not of *EIN3* or *EIL1* (Binder et al., 2004). Thus, non-transcriptional responses to ethylene have been postulated to exist and originate from a signaling pathway diverging at the level or downstream of *EIN2* and, therefore, not including *EIN3*/*EIL1*. Although the existence of this parallel pathway was proposed more than 10 years ago (Binder et al., 2004), the mechanistic understanding of such signaling process has remained obscure. Our finding that ethylene alters the TE of specific genes provided missing evidence to start to uncover the molecular nature of the postulated parallel pathway. The detailed characterization of the ethylene-mediated translational regulation of *EBF2* has revealed that this non-transcriptional ethylene effect was indeed *EIN3*/*EIL1* independent and *EIN2* dependent. Hence, the ethylene-triggered changes in translation fulfilled all the pre-requisites of a long-anticipated branch of the ethylene-signaling pathway diverging at the *EIN2* level. Furthermore, we were able to show that this translation-based signaling branch plays a significant physiological role in the ethylene response. For example, removal of *3′EBF2* resulted in the loss of translational responsiveness of this gene to ethylene, and consequently, dramatic alterations of the plant response to this hormone. In particular, our results indicate that the translational regulation of *EBF2* by ethylene plays a role in the still poorly understood process of plant recovery upon withdrawal of the hormone.

Our results also implicate the key signaling component *EIN2* in the translational regulation of gene expression in response to ethylene. Previous studies have shown that EIN2C moves to the nucleus in the presence of ethylene and that this translocation is required for the activation of the *EIN3*/*EIL1*-dependent transcriptional changes (Qiao et al., 2012). Here, we have shown that EIN2 must also function in a cytosolic process of translational control. Although EIN2 has been implicated in the plant response to a variety of stimuli (Gazzarrini and McCourt, 2003), conclusive evidence for an EIN2 role beyond ethylene signaling is still missing. The finding that EIN2C regulates translation opens new opportunities to investigate the full functional spectrum of this enigmatic protein. An additional mechanism can now be envisioned by which other signals impinge on ethylene signaling—i.e., by altering the translational regulatory activity of EIN2. We also found that the EIN2C localizes to cytoplasmic P-bodies under certain circumstances, such as in *ein5* mutants lacking the 5′-3′ XRN4 exoribonuclease activity (Potuschak et al., 2006; Souret et al., 2004). The observation that EIN2C is retained in the cytosol of *ein5* protoplasts suggests a possible mechanistic explanation for the ethylene insensitivity of this classical ethylene signaling mutant.

In addition to uncovering the EIN2C accumulation in P-bodies, we also showed that EIN2C has the capability to interact, directly or indirectly, with the *3′EBF2* mRNA, as also suggested by the results from the accompanying paper by Li et al. (2015) in this issue of *Cell*. In either case, these results, together with the finding that the *EIN2* function is required for the translational regulation of *EBF2*, raised the question of how EIN2 influences translation activity of its RNA targets. It is possible that the 3′ UTR-bound EIN2C directly or indirectly modulates the activity

(C and D) (C) Representative image and (D) GFP fluorescence of multiple seedlings (n = 7) expressed as the percentage of fluorescence in ACC compared to that in the MS controls. Error bars represent means ± SD. (a) indicates a significant effect of ethylene on the levels of fluorescence (t test, $p < 0.05$).
(E and F) TE of *EBF2* (E) and *EBF1* (F) mRNA, calculated as the relative expression in polysomal/total RNA fractions, in 3-day-old etiolated Col-0 and *upf2-10* seedlings grown in air (Air) or treated with 10 ppm of ethylene for the last 4 hr of the experiment (Ethylene).
The asterisk (*) indicates a significant difference of the ethylene effect on the *EBF* TE between Col and *upf2-10* (two-way ANOVA, $p < 0.05$). (a) indicates a significant difference of the ethylene effect on the *EBF* TE in the indicated genotypes (t test, $p < 0.05$).
The Col measurements in (E) and (F) are the same as in Figure 1D, plotted here again to facilitate the comparison between Col and *upf2-10*. Expression levels of the *EBF* transgenes were normalized against *At4g34270*.

of a component of the general translational machinery, thus selectively inhibiting the translation of its targets. In fact, some of the best-documented examples of gene-specific translation regulation involve the direct interaction of an RNA-binding protein with particular 3′ UTR sequences and a subsequent recruitment of general translation regulators (Szostak and Gebauer, 2013). For example, the *Drosophila* Bicoid protein directly binds to the 3′ UTR of the embryo-patterning mRNA *caudal*. This interaction, however, is not sufficient to repress the translation of *caudal*, and Bicoid has to recruit the CAP-binding protein 4EHP that (due to its low affinity for the translation initiation factor eIF4G) attenuates the rates of translation of the Bicoid targets by failing to recruit the eIF3-containing 43S translation initiation complex (Cho et al., 2005). It is interesting to note here that UPF1 has also been shown to repress translation initiation by directly interacting with eIF3 and, thus, to prevent the formation of the 43S translation initiation complex (Isken et al., 2008). We have provided experimental evidences linking *UPF* function not only with the ethylene response, but also, more specifically, with the translational repression of *EBF2* by this hormone. Furthermore, we show that, under certain experimental conditions, such as in plants lacking functional *EIN5*, the *3′EBF2*-binding protein EIN2C and UPF1 co-localize in P-bodies. Based on this, we propose a mechanistic model (Figure S7) in which the binding of EIN2C to *3′EBF2* triggers the recruitment of the UPFs to this mRNA, which in turn results in the inhibition of translation initiation by interfering with the formation of the 43S complex. Although our initial attempts to show a direct interaction between EIN2C and the UPFs by means of the yeast two hybrid have failed, it is still possible that EIN2C directly or indirectly recruits the UPFs, perhaps, as it has been suggested by Li et al. (2015) in the accompanying paper, via a yet-uncharacterized RNA-binding protein that recruits EIN2 to its target mRNAs. It is also important to point out that the relatively weak ethylene defects observed in the *upf* mutants are likely the result of the hypomorphic nature of the alleles identified in our screen, as well as the fact that the function of UPFs is required for the translational effect of EIN2C but not necessarily for its activation of the EIN3/EIL1 activity. We have focused here on the regulation of *EBF2*, but it would be interesting to study other translationally regulated genes identified herein and explore their role in ethylene-related processes, including transcription-independent fast growth inhibition response (Binder et al., 2004). Finally, additional studies on the temporal dynamics of the transition of the translationally regulated mRNAs from polysomes to P-bodies in ethylene and back to polysomes upon ethylene withdrawal will be necessary to extend the mostly static single-time-point studies described herein.

EXPERIMENTAL PROCEDURES

Plant Growth and Ribosome Footprinting

Plant growth conditions and hormonal treatments of *Arabidopsis* seedlings were as described (Stepanova et al., 2005). Ribosome footprinting (Ingolia et al., 2009) was carried out using pelleted polysomes (Mustroph et al., 2009) with the following modifications. Polysomes were isolated in Extraction Buffer (100 mM Tris-HCl [pH 9], 10 mM Tris-HCl [pH 7.4], 100 mM sucrose, 100 mM KCl, 75 mM NaCl, 20 mM $MgCl_2$, 12.5 mM EGTA [pH8], 3 mM DTT, 6.25 μl/ml detergent mix [20% (w/v or v/v) of each of the four detergents in water: Brij-35, Triton X-100, Igepal CA 630 and Tween 20], 25 μl/ml Triton X-100, 37.5 μg/ml cycloheximide, 25 μg/ml chloramphenicol), and the digestion with the RNase I was carried out in a volume of 4.5 ml. After digestion, monosomes were re-pelleted and purified by sucrose gradient fractionation. RNA fragments corresponding to the ribosome footprints were recovered from the purified monosomes and sequenced as described (Ingolia et al., 2009). Data processing was performed using a combination of custom-made Perl scripts, as well as R and Bioconductor programs.

Immunoblot and qRT-PCR

Protein samples were prepared by homogenizing the liquid nitrogen-ground tissues in 2× SDS-PAGE sample buffer (Laemmli, 1970) and boiling the homogenate for 5 min. Proteins were separated through a 12% SDS-PAGE gel, transferred to a nitrocellulose membrane, and hybridized to anti-GFP antibodies (Living Colors A.v. Monoclonal Antibody, Clontech).

Total RNA was extracted as previously described (Reuber and Ausubel, 1996). Polysomal RNA was isolated by pelleting polysomes (Mustroph et al., 2009) and then extracting the RNA by the SDS/acid phenol method (Ingolia et al., 2009). Reverse transcription and qPCR (Applied Biosystems) were performed according to manufacturer's recommendations. Primer sequences are listed in the Supplemental Experimental Procedures.

Yeast Three-Hybrid and RNA Immunoprecipitation

The yeast three-hybrid system (Bernstein et al., 2002) was used to test the interaction between the EIN2C fragment (amino acids 459 to 1278) and *3′EBF2* RNA. Interaction was inferred based on the activity of LacZ and HIS3 reporters as described (Deplancke et al., 2006).

RNA immunoprecipitation assay was performed as described (Nicaise et al., 2013). Protein extracts from *Nicotiana benthamiana* leaves expressing *35S:GFP-EIN2C-pGWB6* or a negative control *35S:GFP-PDC2-pGWB6* (Stepanova et al., 2011) were incubated with anti-GFP-TRAP-A beads (Chromotek) and 50 μg of *3′EBF2* RNA synthesized in vitro using RiboMAX Large Scale RNA Production System-T7 (Promega). After extensive washes, RNA-protein complexes were eluted from the beads by incubating at 60°C for 15 min in 200 μl of Elution Buffer (1% SDS, 0.1 M NaHCO3) and treated for 1 hr at 60°C with 40 μg Proteinase K, followed by SDS/Phenol RNA extraction, reverse transcription (Applied Biosystems), and 30 cycles of qPCR (Power SYBR green Master Mix, Applied Biosystems).

Protoplast and Tobacco Transient Expression Assays

Protoplasts were isolated using the tape-*Arabidopsis* sandwich method (Wu et al., 2009) and transfected according to a published protocol (Yoo et al., 2007). Transient expression in *Nicotiana benthamiana* leaves was performed as described elsewhere (Wang et al., 2015).

Imaging was done using a Leica DFC365 FX camera attached to a compound microscope DM5000 with the following filters: GFP filter cube (EX 470/40 EM 525/50), CFP filter cube (Ex 436/20 Em 480/40), and TX2 filter cube (Ex 560/40 Em BP645/75). The Objective HCX PLAPO 40×/0.10 was used.

A more detailed description of the materials and methods is provided in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the sequencing data reported in this paper is NCBI SRA: SRP056795.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.036.

AUTHOR CONTRIBUTIONS

C.M., A.N.S., and J.M.A. designed and carried out the experiments and wrote the manuscript. Q.H. and S.H. performed the bioinformatic analysis. B.M.B.

did the kinetic analysis of the ethylene responses. J.B. performed qRT-PCR, yeast, protoplast, and tobacco studies. K.R.S., P.E., and J.Y. assisted in the identification and/or cloning of the mutants.

ACKNOWLEDGMENTS

We thank Hongwei Guo for sharing unpublished data, Valerie Nicaise for technical advice on RNA-IP, Miguel A. Perez-Amador for stimulating discussions on RNA-protein interactions, and Xuemei Chen for *dcl2-1 dcl3-1 dcl4-2*. This work was supported by NSF grants MCB 1158181 and 0519869 to J.M.A.; MCB 0923727 to J.M.A. and A.N.S.; IOS 1444561 to J.M.A., A.N.S., and S.H.; NCSU-RISF to S.H. and J.M.A.; a Marie Curie COFUND U-Mobility postdoctoral fellowship to C.M. (co-funded by the University of Málaga and the EU 7FP GA N°246550); a postdoctoral fellowship from Ministerio de Educacion 2008-2011 to J.B.; and an NSF-REU MCB 06103224 to K.R.S. Sequencing data have been deposited at the NCBI SRA (SRP056795).

Received: April 2, 2015
Revised: August 4, 2015
Accepted: September 8, 2015
Published: October 22, 2015

REFERENCES

Abeles, F., Morgan, P., and Saltveit, M.J. (1992). Ethylene in Plant Biology (Academic Press).

Alonso, J.M., Hirayama, T., Roman, G., Nourizadeh, S., and Ecker, J.R. (1999). EIN2, a bifunctional transducer of ethylene and stress responses in Arabidopsis. Science *284*, 2148–2152.

Alves, L., Jr., Niemeier, S., Hauenschild, A., Rehmsmeier, M., and Merkle, T. (2009). Comprehensive prediction of novel microRNA targets in Arabidopsis thaliana. Nucleic Acids Res. *37*, 4010–4021.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

Bernstein, D.S., Buter, N., Stumpf, C., and Wickens, M. (2002). Analyzing mRNA-protein complexes using a yeast three-hybrid system. Methods *26*, 123–141.

Binder, B.M., Mortimore, L.A., Stepanova, A.N., Ecker, J.R., and Bleecker, A.B. (2004). Short-term growth responses to ethylene in Arabidopsis seedlings are EIN3/EIL1 independent. Plant Physiol. *136*, 2921–2927.

Binder, B.M., Walker, J.M., Gagne, J.M., Emborg, T.J., Hemmann, G., Bleecker, A.B., and Vierstra, R.D. (2007). The Arabidopsis EIN3 binding F-Box proteins EBF1 and EBF2 have distinct but overlapping roles in ethylene signaling. Plant Cell *19*, 509–523.

Bleecker, A.B., Estelle, M.A., Somerville, C., and Kende, H. (1988). Insensitivity to ethylene conferred by a dominant mutation in Arabidopsis thaliana. Science *241*, 1086–1089.

Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.S., Schmitz, R.J., Urich, M.A., Kuo, D., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife *2*, e00675.

Cho, P.F., Poulin, F., Cho-Park, Y.A., Cho-Park, I.B., Chicoine, J.D., Lasko, P., and Sonenberg, N. (2005). A new paradigm for translational control: inhibition via 5′-3′ mRNA tethering by Bicoid and the eIF4E cognate 4EHP. Cell *121*, 411–423.

Clark, K.L., Larsen, P.B., Wang, X., and Chang, C. (1998). Association of the Arabidopsis CTR1 Raf-like kinase with the ETR1 and ERS ethylene receptors. Proc. Natl. Acad. Sci. USA *95*, 5401–5406.

de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Deplancke, B., Vermeirssen, V., Arda, H.E., Martinez, N.J., and Walhout, A.J. (2006). Gateway-compatible yeast one-hybrid screens. CSH Protoc. *2006*, pdb.prot4590.

Gazzarrini, S., and McCourt, P. (2003). Cross-talk in plant hormone signalling: what Arabidopsis mutants are telling us. Ann. Bot. (Lond.) *91*, 605–612.

Goeres, D.C., Van Norman, J.M., Zhang, W., Fauver, N.A., Spencer, M.L., and Sieburth, L.E. (2007). Components of the Arabidopsis mRNA decapping complex are required for early seedling development. Plant Cell *19*, 1549–1564.

Guo, H., and Ecker, J.R. (2003). Plant responses to ethylene gas are mediated by SCF(EBF1/EBF2)-dependent proteolysis of EIN3 transcription factor. Cell *115*, 667–677.

Guzmán, P., and Ecker, J.R. (1990). Exploiting the triple response of Arabidopsis to identify ethylene-related mutants. Plant Cell *2*, 513–523.

He, W., Brumos, J., Li, H., Ji, Y., Ke, M., Gong, X., Zeng, Q., Li, W., Zhang, X., An, F., et al. (2011). A small-molecule screen identifies L-kynurenine as a competitive inhibitor of TAA1/TAR activity in ethylene-directed auxin biosynthesis and root growth in Arabidopsis. Plant Cell *23*, 3944–3960.

Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nat. Genet. *38*, 721–725.

Hua, J., and Meyerowitz, E.M. (1998). Ethylene responses are negatively regulated by a receptor gene family in Arabidopsis thaliana. Cell *94*, 261–271.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Isken, O., Kim, Y.K., Hosoda, N., Mayeur, G.L., Hershey, J.W., and Maquat, L.E. (2008). Upf1 phosphorylation triggers translational repression during nonsense-mediated mRNA decay. Cell *133*, 314–327.

Ju, C., Yoon, G.M., Shemansky, J.M., Lin, D.Y., Ying, Z.I., Chang, J., Garrett, W.M., Kessenbrock, M., Groth, G., Tucker, M.L., et al. (2012). CTR1 phosphorylates the central regulator EIN2 to control ethylene hormone signaling from the ER membrane to the nucleus in Arabidopsis. Proc. Natl. Acad. Sci. USA *109*, 19486–19491.

Konishi, M., and Yanagisawa, S. (2008). Ethylene signaling in Arabidopsis involves feedback regulation via the elaborate control of EBF2 expression by EIN3. Plant J. *55*, 821–831.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680–685.

Li, W., Ma, M., Feng, Y., Li, H., Wang, Y., Ma, Y., Li, M., An, F., and Guo, H. (2015). EIN2-directed translational regulation of ethylene signaling in Arabidopsis. Cell *163*, this issue, 670–683.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

Mustroph, A., Juntawong, P., and Bailey-Serres, J. (2009). Isolation of plant polysomal mRNA by differential centrifugation and ribosome immunopurification methods. Methods Mol. Biol. *553*, 109–126.

Nicaise, V., Joe, A., Jeong, B.R., Korneli, C., Boutrot, F., Westedt, I., Staiger, D., Alfano, J.R., and Zipfel, C. (2013). Pseudomonas HopU1 modulates plant immune receptor levels by blocking the interaction of their mRNAs with GRP7. EMBO J. *32*, 701–712.

Olmedo, G., Guo, H., Gregory, B.D., Nourizadeh, S.D., Aguilar-Henonin, L., Li, H., An, F., Guzman, P., and Ecker, J.R. (2006). ETHYLENE-INSENSITIVE5 encodes a 5′–>3′ exoribonuclease required for regulation of the EIN3-targeting F-box proteins EBF1/2. Proc. Natl. Acad. Sci. USA *103*, 13286–13293.

Pomeranz, M.C., Hah, C., Lin, P.C., Kang, S.G., Finer, J.J., Blackshear, P.J., and Jang, J.C. (2010). The Arabidopsis tandem zinc finger protein AtTZF1 traffics between the nucleus and cytoplasmic foci and binds both DNA and RNA. Plant Physiol. *152*, 151–165.

Potuschak, T., Lechner, E., Parmentier, Y., Yanagisawa, S., Grava, S., Koncz, C., and Genschik, P. (2003). EIN3-dependent regulation of plant ethylene hormone signaling by two arabidopsis F box proteins: EBF1 and EBF2. Cell *115*, 679–689.

Potuschak, T., Vansiri, A., Binder, B.M., Lechner, E., Vierstra, R.D., and Genschik, P. (2006). The exoribonuclease XRN4 is a component of the ethylene response pathway in Arabidopsis. Plant Cell *18*, 3047–3057.

Qiao, H., Chang, K.N., Yazaki, J., and Ecker, J.R. (2009). Interplay between ethylene, ETP1/ETP2 F-box proteins, and degradation of EIN2 triggers ethylene responses in Arabidopsis. Genes Dev. *23*, 512–521.

Qiao, H., Shen, Z., Huang, S.S., Schmitz, R.J., Urich, M.A., Briggs, S.P., and Ecker, J.R. (2012). Processing and subcellular trafficking of ER-tethered EIN2 control response to ethylene gas. Science *338*, 390–393.

Reuber, T.L., and Ausubel, F.M. (1996). Isolation of Arabidopsis genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. Plant Cell *8*, 241–249.

SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., and Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. Proc. Natl. Acad. Sci. USA *93*, 8496–8501.

Souret, F.F., Kastenmayer, J.P., and Green, P.J. (2004). AtXRN4 degrades mRNA in Arabidopsis and its substrates include selected miRNA targets. Mol. Cell *15*, 173–183.

Stepanova, A.N., Hoyt, J.M., Hamilton, A.A., and Alonso, J.M. (2005). A Link between ethylene and auxin uncovered by the characterization of two root-specific ethylene-insensitive mutants in Arabidopsis. Plant Cell *17*, 2230–2242.

Stepanova, A.N., Yun, J., Robles, L.M., Novak, O., He, W., Guo, H., Ljung, K., and Alonso, J.M. (2011). The Arabidopsis YUCCA1 flavin monooxygenase functions in the indole-3-pyruvic acid branch of auxin biosynthesis. Plant Cell *23*, 3961–3973.

Szostak, E., and Gebauer, F. (2013). Translational control by 3′-UTR-binding proteins. Brief. Funct. Genomics *12*, 58–65.

Vandenbussche, F., Vaseva, I., Vissenberg, K., and Van Der Straeten, D. (2012). Ethylene in vegetative development: a tale with a riddle. New Phytol. *194*, 895–909.

Wang, G.F., Ji, J., El-Kasmi, F., Dangl, J.L., Johal, G., and Balint-Kurti, P.J. (2015). Molecular and functional analyses of a maize autoactive NB-LRR protein identify precise structural requirements for activity. PLoS Pathog. *11*, e1004674.

Weber, C., Nover, L., and Fauth, M. (2008). Plant stress granules and mRNA processing bodies are distinct from heat stress granules. Plant J. *56*, 517–530.

Wu, F.H., Shen, S.C., Lee, L.Y., Lee, S.H., Chan, M.T., and Lin, C.S. (2009). Tape-Arabidopsis Sandwich - a simpler Arabidopsis protoplast isolation method. Plant Methods *5*, 16.

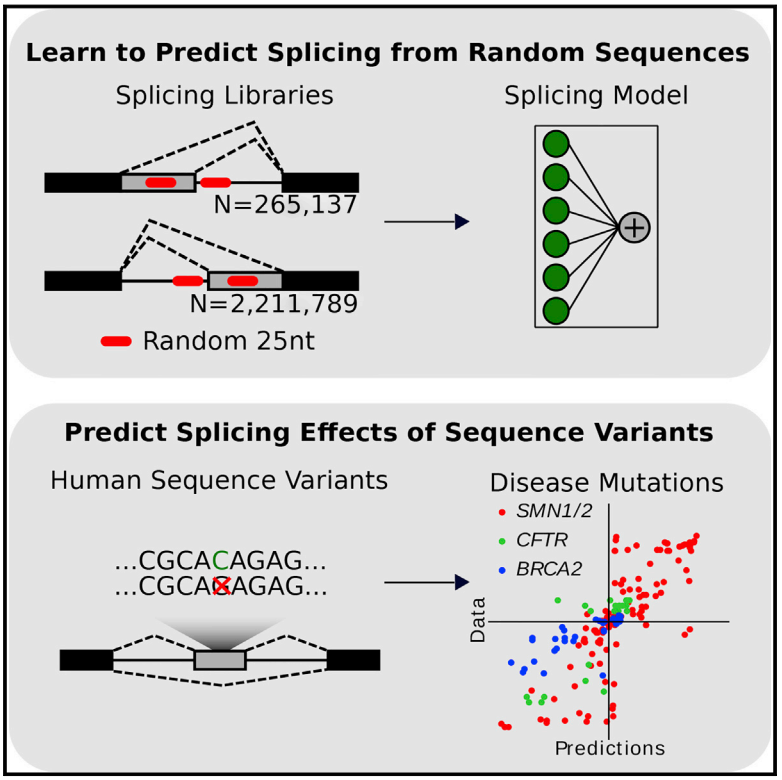
Yoo, S.D., Cho, Y.H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat. Protoc. *2*, 1565–1572.

Article

Cell

Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences

Graphical Abstract



Authors

Alexander B. Rosenberg, Rupali P. Patwardhan, Jay Shendure, Georg Seelig

Correspondence

gseelig@uw.edu

In Brief

A combination of synthetic biology and machine learning approaches identifies universal rules of RNA splicing and enables the accurate prediction of the effects of disease-related human SNPs on isoform levels.

Highlights

- Measured splicing patterns of nearly 2M synthetic alternatively spliced mini-genes
- *cis*-regulatory elements primarily act additively rather than cooperatively
- Model trained only on synthetic data predicts effects of human SNPs on isoform ratios
- Model of alternative 5′ and 3′ splicing predicts effect of SNPs in skipped exons

Accession Numbers

GSE74070

Rosenberg et al., 2015, Cell *163*, 698–711
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.054

CellPress

Cell

Article

Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences

Alexander B. Rosenberg,[1] Rupali P. Patwardhan,[2] Jay Shendure,[2] and Georg Seelig[1,3,*]
[1]Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
*Correspondence: gseelig@uw.edu
http://dx.doi.org/10.1016/j.cell.2015.09.054

SUMMARY

Most human transcripts are alternatively spliced, and many disease-causing mutations affect RNA splicing. Toward better modeling the sequence determinants of alternative splicing, we measured the splicing patterns of over two million (M) synthetic mini-genes, which include degenerate subsequences totaling over 100 M bases of variation. The massive size of these training data allowed us to improve upon current models of splicing, as well as to gain new mechanistic insights. Our results show that the vast majority of hexamer sequence motifs measurably influence splice site selection when positioned within alternative exons, with multiple motifs acting additively rather than cooperatively. Intriguingly, motifs that enhance (suppress) exon inclusion in alternative 5′ splicing also enhance (suppress) exon inclusion in alternative 3′ or cassette exon splicing, suggesting a universal mechanism for alternative exon recognition. Finally, our empirically trained models are highly predictive of the effects of naturally occurring variants on alternative splicing in vivo.

INTRODUCTION

Alternative splicing is a major source of proteome diversity in eukaryotes (Nilsen and Graveley, 2010). Regulation of alternative splicing is vital to cellular processes that depend on the precise ratios of isoforms. For example, mutations that lead to even subtle changes in the ratio of *MAPT* isoforms 3R and 4R cause an inherited form of dementia (Garcia-Blanco et al., 2004). While new sequencing technologies have enabled the comprehensive cataloging of human genetic variation, the functional consequences of these variants on even molecular phenotypes, such as alternative splicing, remain poorly predictable.

Experimentally testing the consequence of every possible genetic variant on endogenous alternative splicing is impractical, motivating the development of predictive models of the "splicing code." The core splicing signals—5′ splice donor, 3′ splice acceptor, branchpoint, and polypyrimidine tract—form the basis of the splicing code; they are required for recognition of intron-exon boundaries and for correct intron removal by the splicing machinery. Computational methods have been developed to score the likelihood of splicing at different splice donor and acceptor sequences (Yeo and Burge, 2004). Splice regulatory elements (SREs)—sequence motifs in exons or introns shown to regulate splicing—form the next level of regulatory information. SREs typically regulate alternative splicing by binding *trans*-acting splice factor proteins (Ule et al., 2006; Wang et al., 2013). Depending on their position and mode of action, SREs are classified as exonic splice enhancers (ESEs), exonic splice silencers (ESSs), intronic splice enhancers (ISEs), or intronic splice silencers (ISSs). Examples of SREs have been identified computationally by analyzing motif enrichment near splice sites (Castle et al., 2008; Fairbrother et al., 2002; Zhang and Chasin, 2004) or sequence conservation between species (Goren et al., 2006). Recently, a deep neural network was trained on exon skipping events in the genome to generate a comprehensive model of the splicing code that can be used to predict exon inclusion percentages (Xiong et al., 2014). Despite this progress, current models of alternative splicing do not perform well enough to be used in clinical genetics (e.g., to reclassify "variants of uncertain significance"), and many machine learning strategies result in "black boxes" that limit mechanistic insight.

We hypothesized that a model of alternative splicing learned from very large libraries of synthetic sequences could outperform models trained only on the genome. Current technology makes it possible to create and test gene libraries with millions of synthetic sequences—orders of magnitude more than the number of alternative splice events in the human genome. In other applications of machine learning, such as computer vision, predictive power has increased greatly with access to larger datasets (Le et al., 2012). Previous work supports the idea that synthetic gene libraries with extensive and targeted variation can provide mechanistic insight into biological phenomena. In vivo (Culler et al., 2010; Wang et al., 2012) and in vitro (Yu et al., 2008) randomized selections have identified potential SREs. Massively parallel reporter assays (MPRAs) that combine next-generation sequencing with extensive variation have been applied to study transcription (Melnikov et al., 2012; Patwardhan et al., 2012; Patwardhan et al., 2009; Sharon et al., 2012; Smith et al., 2013; White et al., 2013), translation (Noderer et al., 2014), mRNA stability (Oikonomou et al., 2014), and even alternative

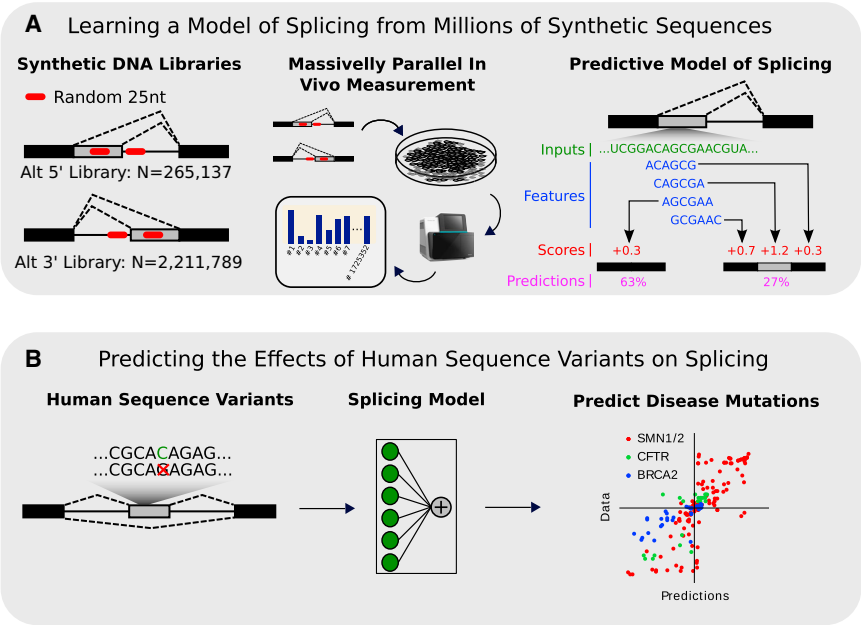


Figure 1. A Predictive Model of Alternative Splicing Learned from Millions of Synthetic Sequences

(A) Two libraries with either alternative 5′ or 3′ splice sites were constructed with two 25-nt randomized regions. The library was transfected into human cells, and massively parallel measurement of isoform ratios was performed with RNA-seq. These two datasets were used to learn a predictive model of alternative splicing. The model takes a sequence as input, which is then converted to 6-mer features. A score for each 6-mer is learned and then used to predict the fractional usage of each splice site.

(B) When human sequence variants are fed to the model as inputs, the model makes more accurate predictions than the current state of the art algorithms.

splicing (Ke et al., 2011). However, MPRA studies to date have overwhelmingly focused on measuring the consequences of variants in endogenous sequences (e.g., saturation mutagenesis) or on validating predicted activities (e.g., enhancers predicted by the ENCODE project). There are thus far few, if any, examples of predictive biological models learned entirely on MPRA data.

To test whether it is possible to learn predictive biological models from synthetic data alone, we developed an MPRA that measures alternative splice site selection in a highly complex library of "degenerate introns" (Figure 1A). We added degenerate regions into an otherwise fixed sequence context, ensuring that any differences in gene expression can be causally attributed to the degenerate region. We created two libraries, one with alternative 5′ splice donors consisting of 265,137 members and one with alternative 3′ splice acceptors containing 2,211,739 members. We transfected these libraries to human cells, performed RT-PCR and RNA sequencing (RNA-seq) to quantitatively measure isoform ratio for all mini-genes and used the results to learn a predictive model of alternative splicing. To assess the quality of the resulting model, we predicted the effects of human sequence variants on isoform levels and compared our results to available experimental data (Figure 1B). We tested variants in alternative 5′ splicing events, both within the alternative splice donors themselves and within the alternative exon. Although our MPRA did not include a skipped exon library, our model also predicted with high accuracy the effect of sequence variants in skipped exons.

RESULTS

Molecular Phenotyping of Millions of Alternatively Spliced Mini-Genes Containing Random Sequences

We chose to study both alternative 5′ and alternative 3′ splice site selection. In the case of alternative 5′ splicing, we first generated a complex library by introducing 2 × 25 nt fully degenerate regions into a single-intron plasmid mini-gene (Figure 2A). Specifically, the intron was designed with two competing splice donors separated by 44 nt; one degenerate region was inserted between the splice donors and the other downstream of the second donor. Neither degenerate sequence overlapped a splice donor. The mini-genes contained an additional degenerate 20 nt barcode in the 3′ UTR. This barcode was used to create a look-up table linking barcodes and intronic sequences. Thus, even when both degenerate regions were spliced out, their sequences could be recovered from the barcode sequence (Figure 2A). To maximize intron sequence variability, we constructed and sequenced a complex library of 265,137 such mini-genes. Thus, over 13 Mb of unique intronic sequence are represented within the degenerate regions of this library (265,137 × 50 nt).

In the case of alternative 3′ splicing, we inserted 2 × 25 nt fully degenerate regions into a single-intron system designed to have two alternative 3′ splice sites (Figure 2C). The degenerate regions did not overlap either splice acceptor, but the upstream degenerate region did overlap the typical position of the first splice acceptor's branchpoint (−44:−19 relative to SA_1). Similarly to the alternative 5′ library, we included an additional degenerate 20-nt barcode in the 3′ UTR. The alternative 3′ library contained 2.2 million unique mini-genes encompassing over 110 Mb of unique sequence variation (2,211,739 × 50 nt).

We transfected the pooled libraries of plasmids into HEK293 cells and then quantified isoform ratios with targeted RNA-seq. To identify both the isoform and originating plasmid of each mRNA, we used paired-end sequencing with one read across the exon junction and the other read across the 3′ UTR barcode (Figures 2A and 2C). We used 13 million reads for the alternative 5′ library and 5.4 million reads for the alternative 3′ library. We were then able to calculate the isoform ratios for each mini-gene in each library. We averaged 50.0 reads per mini-gene in the 5′ library with reads mapping to 265,044/265,137 (99.96%) of all mini-genes. On the other hand, in the 3′ library we averaged

only 2.47 reads per mini-gene with reads mapping to 1,686,096/2,211,739 (76.23%) of all mini-genes.

Degenerate Sequences in Both Libraries Strongly Influence Isoform Ratios

In the alternative 5′ library, isoforms were present from several different splicing events. The most upstream splice donor (SD_1) was used on average 22.4% of the time, while SD_2 was used 50.0% of the time (Figure 2B). The remaining transcripts were spliced at new splice donors inserted into the randomized regions (11.3%), a cryptic splice donor site (SD_{CRYPT}) 35 nt downstream of SD_2 (7.9%), or not spliced at all (8.4%). However, as evidenced by the broad distributions of usage at each SD (Figure 2B), the degenerate regions had a strong influence on splice site selection. For instance, although 49.7% of mini-genes spliced at SD_1 with less than 5% frequency, 7,705 mini-genes (2.9%) spliced at SD_1 with over 95% frequency.

In the alternative 3′ library, we also found isoforms from different splicing events, although splice site usage was less evenly balanced than in the 5′ library. SA_1 was used an average of 3.3% of the time, while SA_2 was used 89.2% of the time (Figure 2D). In this library, new splice sites in the randomized regions were only used with 0.3% frequency, probably reflecting the larger information footprint of splice acceptors (>20 nt) compared to splice donors (9 nt), which makes the occurrence of new sites within the degenerate regions less likely. Similarly to the 5′ library, we inadvertently inserted a cryptic splice acceptor 16 nt upstream of SA_2 that was used with 4.6% frequency. Many other cryptic splice sites were used with very low frequency (1×10^{-7} to 5×10^{-3}) accounting for a total of 2.3% of transcripts. In contrast with the alternative 5′ library, only 0.3% of transcripts were unspliced. Although SA_2 was the dominant splice site, 0.7% of the 1.2 M of mini-genes represented by multiple reads spliced 100% at SA_1.

With so many transcripts in each library splicing at new splice sites, we asked whether we could rediscover the known motifs for splice donors and splice acceptors from the de novo sites alone. When we plotted the relative frequencies of each base at each position for new splice donors (Figure 2E) and new splice acceptors (Figure 2F), both splice site motifs were nearly identical to the expected motifs for splice donors and splice acceptors. More specifically, the splice donors contained the canonical GT at the +1:+2 positions, while the splice acceptors contain a clear polypyrimidine tract (T and C rich), followed by N[CT]AGG. The ability to fully rediscover canonical signals for splice donors and splice acceptors demonstrates the rich type of information contained in each dataset.

We also asked whether translation might affect the mRNA stability in our libraries. Sequencing of the alternative 5′ library yielded fewer median reads on mRNA from mini-genes that were primarily spliced out of frame than in frame (Figure S1A). However, when the mini-genes contained a premature stop codon, the median number of reads per mRNA was similar for all three reading frames (Figure S1B). These results indicate that a large string of amino acids translated out of frame will destabilize the mRNA, likely through the no-go decay pathway (Doma and Parker, 2006; Shoemaker et al., 2010) as ribosomes stall due to protein misfolding. We also find evidence of nonsense-mediated decay, but only if the premature stop codon occurred >40 nt upstream of the splice donor. This is consistent with previous studies on nonsense-mediated decay that suggest the premature stop codon must occur >50 nt upstream of the last exon junction (Lewis et al., 2003).

Splicing Is More Likely to Occur at Upstream Splice Donors

From an analysis of the new splice sites, we found strong evidence that upstream splice donors were favored over downstream splice donors; new splice donors inserted in the first degenerate region were 4.1 times more likely to be used than new splice donors inserted into the second degenerate region (region 1: 849,666 spliced reads; region 2: 208,396 spliced reads). Furthermore, the effect of position of splice donors within each degenerate region was significant ($p < 0.005$; Figure S1C). The number of spliced reads at a new splice site decayed exponentially with the distance from SD_1 (Figure 2G). Splicing has been shown to be co-transcriptional, and spliceosome components can begin to assemble at a 5′ splice donor before downstream alternative slice sites are transcribed (Listerman et al., 2006), suggesting a potential mechanistic explanation for the observed effect. This strong bias for upstream splice donors is consistent with the typically short length of exons in the human genome (Burge and Karlin, 1997).

Splicing Is Less Likely to Occur at Splice Acceptors with Distal Branchpoints

Large-scale mapping of human branchpoints with RNA-seq found that 90% of mapped branchpoints occur between 19–37 nt upstream of the splice acceptor (Mercer et al., 2015). However, it remains unclear just how detrimental a distal branchpoint is toward efficient splicing. Consensus branchpoints (CU[AG]A[CU]) occur over 10,000 times at every position between 40 to 19 nt upstream of SA_1 in our dataset, allowing us to answer this question. We found that mini-genes with a consensus branchpoint sequence 19 nt upstream of SA_1 were approximately six times more likely to be spliced at SA_1 relative to those with a branchpoint 40 nt upstream of SA_1 (Figure 2H). One explanation for this phenomenon could be that distal branchpoints are more likely to contain another AG between the branchpoint and SA_1 that could be used as an alternative splice acceptor. However, we observed a strong distance dependence on branchpoint position for sequences both with and without an AG between the branchpoint and SA_1 (Figure S1D). This result suggests that mechanism by which distal branchpoints reduce splicing efficiency is primarily due to the increased distance between the branchpoint and the splice acceptor and/or polypyrimidine tract.

Sequence Motifs in Alternative Exons Have a Stronger Regulatory Role than Intronic Sequences

Next, we asked how short sequence motifs affect splice site selection in different contexts. We chose to analyze the effects of 6-mer because each possible 6-mer occurs within an average of 1,294 mini-genes for the alternative 5′ library, and 8,232 mini-genes for the alternative 3′ library. Furthermore, most known RNA binding proteins (RBPs) are reported to bind sequences between 4–8 nt (Lunde et al., 2007). In order to estimate the effect of

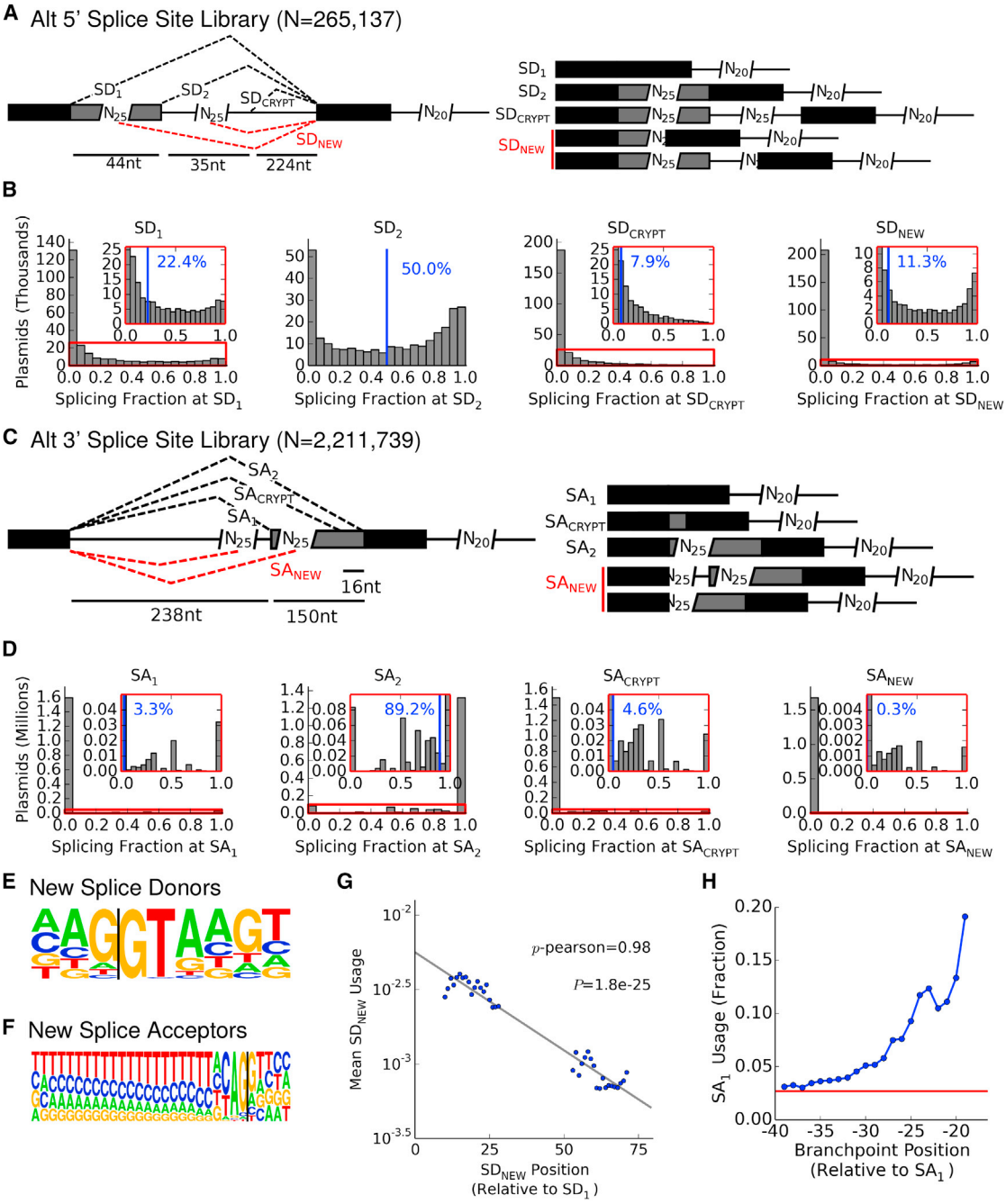


Figure 2. Splice Site Selection in Two Million Alternative 5′ and 3′ Spliced Sequences

(A) A schematic of the alternative 5′ library. Spliced reads map to SD_1, SD_2, and a cryptic splice site (SD_{CRYPT}), as well as new splice donors (SD_{NEW}) created in the degenerate regions.

(B) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.

(C) A schematic of the alternative 3′ library. Spliced reads map to SA_1, SA_2, and a cryptic splice site (SA_{CRYPT}), as well as new splice donors (SD_{NEW}) created in the degenerate regions.

(D) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.

(E) The splice donor motif recovered from the new splice alternative 5′ library matches the previously known human splice donor site.

(F) The splice acceptor motif recovered from the new splice alternative 5′ library matches the previously known human splice acceptor site.

(legend continued on next page)

each possible 6-mer in each region, we calculated splice site usage for the subset of mini-genes containing the 6-mer and for the much larger subset not containing the motif. We then asked to what extent the odds of splicing at a splice site changed in the presence of the motif relative to the control set. To quantify this "effect size," we used the $\log_2$ odds ratio with and without the 6-mer present (Supplemental Experimental Procedures). For example, we found that mini-genes containing the 6-mer GTGGGG in the first degenerate region of the 5′ library were spliced at SD_2 only 19.0% of the time, while RNA derived from mini-genes not containing this motif spliced at SD_2 50.2% of the time, resulting in an effect size of −2.1 (Figures S2A–S2D). In other words, the odds of splicing at SD_2 are 4.29 ($2^{2.1}$) times lower in the presence of GTGGGG compared to its absence.

In Figure 3A, we plot the empirically measured effect sizes of all hexamers in the first degenerate region on the relative usage of SD_2 and SD_1, with 95% confidence intervals. The strongest enhancers located in the alternative exon (included when splicing occurs at SD_2, but excluded when splicing occurs at SD_1) increased the odds of splicing at SD_2 4.38-fold, while the strongest silencers decreased the odds 16-fold. Approximately 15% of 6-mer have been previously identified as SREs (Culler et al., 2010; Fairbrother et al., 2002; Wang et al., 2004, 2012) (622/4,096), but here 82.9% of 6-mer (3,396/4,096) exhibited a significant effect on isoform selection (95% confidence interval does not contain zero effect size). Intriguingly, the cumulative effects of previously identified SREs accounted for only 20% of the cumulative effects of all possible 6-mer. The strongest silencers were G rich, consistent with known binding sites for hnRNPs (Martinez-Contreras et al., 2006). On the other hand, some of the strongest enhancers for SD_2 appear to act by generating secondary structure around SD_1: the 6-mer perfectly complementary to part of SD_1 (−3 to +8) were all in the top 6% of SD_2 enhancers (percentiles: 97.77, 99.75, 99.97, 94.23, 94.79, and 98.92).

We then looked at the effects of 6-mers in the second degenerate region (3′ to SD_2). Unlike the first degenerate region, which is located within the alternative exon region, the second degenerate region is intronic to both SD_1 and SD_2. We found that the effect sizes were much smaller than in the first degenerate region (Figure 3B). The strongest enhancer and silencer of SD_2, respectively, only changed the odds of splicing at SD_2 relative to SD_1 1.95-fold and 1.48-fold. Furthermore, only 36.7% of 6-mer (1,505/4,096) had a statistically significant effect.

We performed a similar analysis for each degenerate region on the usage of SA_1 in the alternative 3′ library (Figures 3C and 3D). Again, we found that motifs in the alternative exon (3′ of SA_1, but 5′ of SA_2) had strong effect sizes (statistically significant 6-mer effect sizes: 3,500/4,096, 85.4%; strongest enhancer: 3.84-fold increase in odds of splicing at SD_2; strongest silencer: 9.87-fold decrease in odds of splicing at SD_2). Unlike in the alternative 5′ library, we found that motifs in the intronic degenerate region (5′ of SA_1 and SA_2) also had quite strong effects (statistically significant 6-mer effect sizes: 3,248/4,096, 79.3%; strongest enhancer: 3.45-fold increase in odds-ratio; strongest silencer: 4.63-fold decrease in odds-ratio), although still generally smaller in magnitude than the downstream alternative exon region. When we looked at the strongest 6-mer enhancers of SA_1 in this intronic region, we found they all fit the consensus branchpoint sequence CU[AG]A[CU] (Figure 3D).

The Same Sequence Motifs Regulate Alternative Exon Inclusion Independent of the Type of Alternative Splicing

Surprisingly, we found that the effect sizes of 6-mers occurring within the alternative exon regions were extremely similar between the alternative 5′ and 3′ libraries (Figure 3E; $R^2 = 0.68$). We looked at several motifs known to bind splice factors or that have previously been identified as ESEs/ESSs (G-run, SRSF1, hnRNPA1, hnRNPH2) and found the effect sizes to be highly correlated. In both libraries, GGGGGG was the strongest exonic silencer (5′ library: 16.0-fold change in odds ratio; 3′ library: 9.87-fold reduction in odds ratio).

We also compared the effect sizes of intronic 6-mers (second randomized region in the alt. 5 library; first randomized region in the alt. 3′ library) between the two libraries. We found a significant, but weaker, correlation between the 6-mer scores ($R^2 = 0.27$; Figure S2E). The first randomized region in the alternative 3′ library overlaps the expected location of the SA_1 branchpoint, which may reduce the effect size correlation. However, the weaker correlation can also be explained by the fact that the effect sizes of intronic 6-mer were much smaller in magnitude compared to 6-mer within the alternative exon regions.

Sequence Motifs Regulate Exon Inclusion Additively Rather than Cooperatively

Although previous studies have observed co-occurrence of conserved sequence motifs around splice sites (Barash et al., 2010), it remains unclear whether such motifs act cooperatively or additively and independently of one another to regulate alternative splicing. In an additive and independent model of regulation, the joint effect size of multiple motifs should simply equal the sum of the individual effect sizes (Figure 4A). To assess this, we examined the joint effect sizes of pairs of 4-mers on alternative exon-inclusion levels in both the 5′ and 3′ libraries. We chose 4-mers because pairs of 4-mers occur sufficiently often within each randomized region to allow for robust effect size measurements (alt. 5′ library: 692 mini-genes/4-mer pair; alt. 5′ library: 4,399 mini-genes/4-mer pair).

We first calculated the individual effect size of all 4-mers on exon inclusion in the 5′ library. We then calculated the joint effect size of every possible pair of non-overlapping 4-mers. Surprisingly, we found that combinatorial effects were extremely well

(G) The number of spliced reads at each position within the randomized regions shows a strong position dependency. Splicing is more likely to occur at an upstream (5′) splice donor than at a downstream (3′) splice donor. The gray line is a fit that shows the linear relationship between the location of splice donor and the log read count at that location.

(H) Mini-genes with a consensus branchpoint near SA_1 are much more likely to use SA_1 than mini-genes with a distal branchpoint. The red line indicates the SA_1 usage, when there is no consensus branchpoint.

See also Figure S1.

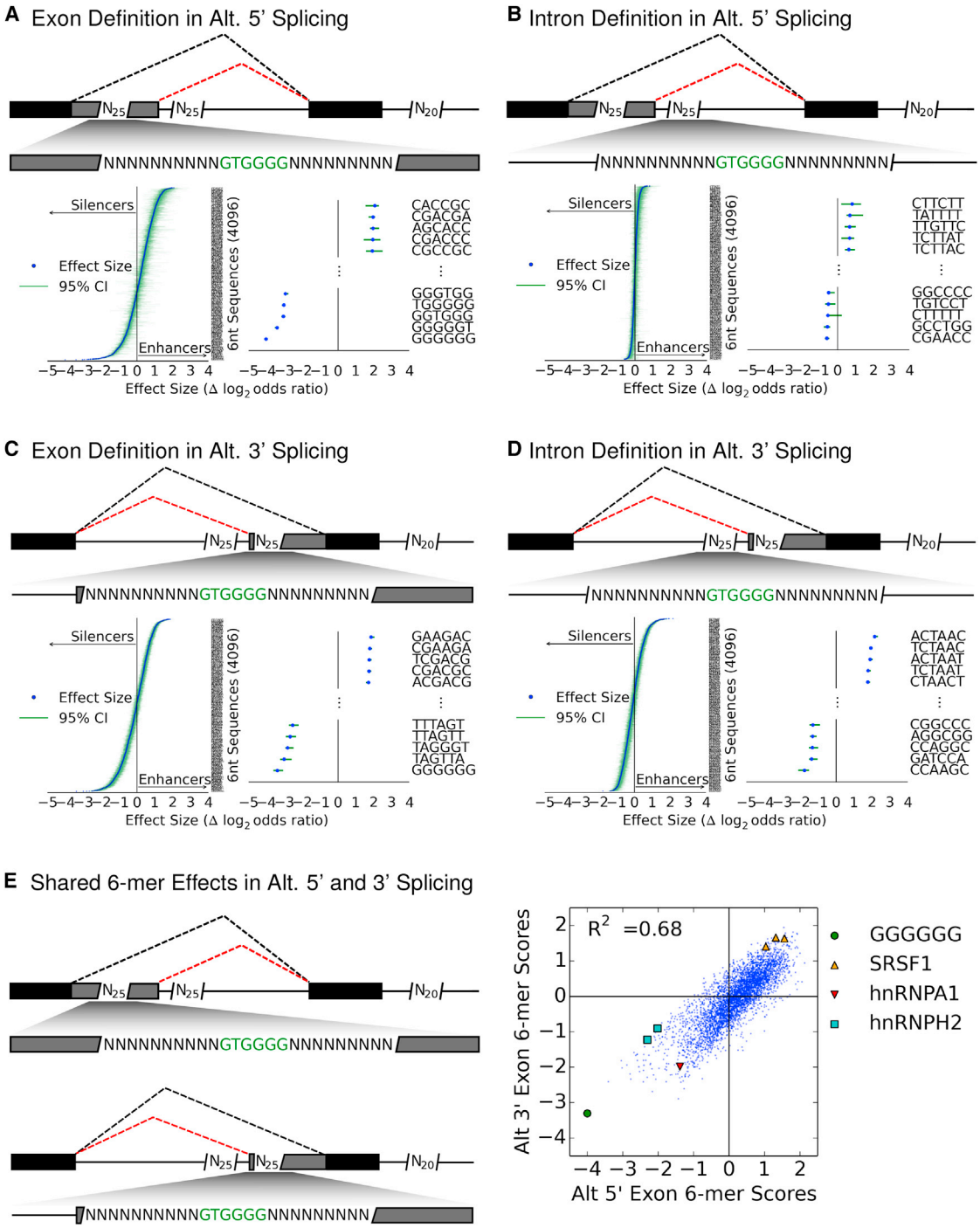


Figure 3. Measured Effect Sizes of Individual 6-mer in Each Degenerate Region

(A–D) To measure how sequence motifs alter the relative use of SD_2/SD_1 or SA_1/SA_2, we calculated effect sizes for every 6-mer (n = 4,096) within each degenerate region in both libraries. We defined effect sizes as the log odds ratio of SD_2 or SA_1 usage between mini-genes with/without the 6-mer of interest. The 6-mer are ranked by estimated effect size and plotted with 95% confidence intervals generated by bootstrapping with replacement. (A) Alternative exon region in 5′ library. (B) Intronic region in 5′ library. (C) Alternative exon region in 3′ library. (D) Intronic region in 3′ library.

(E) The 6-mer scores in the alternative exon region in both the 5′ and 3′ libraries (A and C) are highly similar, suggesting alternative splicing in both libraries is regulated by the same mechanism.

See also Figure S2.

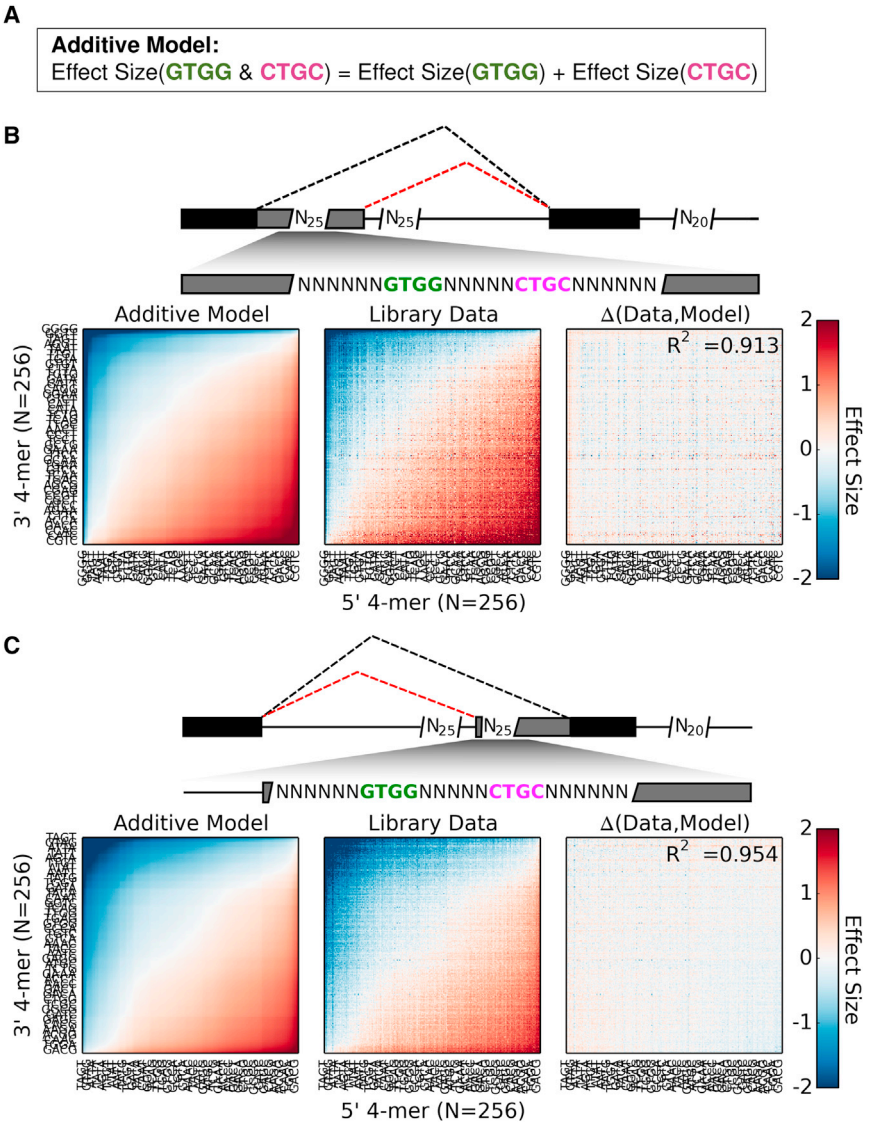


Figure 4. Combinatorial Regulation of Alternative Splicing Is Additive

(A) An additive model of alternative splicing regulation: the joint effect size of two 4-mer is equal to the sum of the individual 4-mer effects.

(B) Using an additive model, the predicted combinatorial effect size of every pair of 4-mer (n = 65,536) is plotted on the left. Each pixel corresponds to a pair of 4-mer with the 5′ 4-mer on the x axis and the 3′ 4-mer on the y axis. The measured combinatorial effect sizes from the 5′ library data are plotted in the middle. The residuals between the additive model and the observed data are plotted on the right. The additive model explains >90% of the combinatorial effect sizes (R^2 = 0.913).

(C) The same analysis is repeated for the alternative 3′ library. In this library the additive model explains over 95% of the combinatorial effect sizes (R^2 = 0.954).

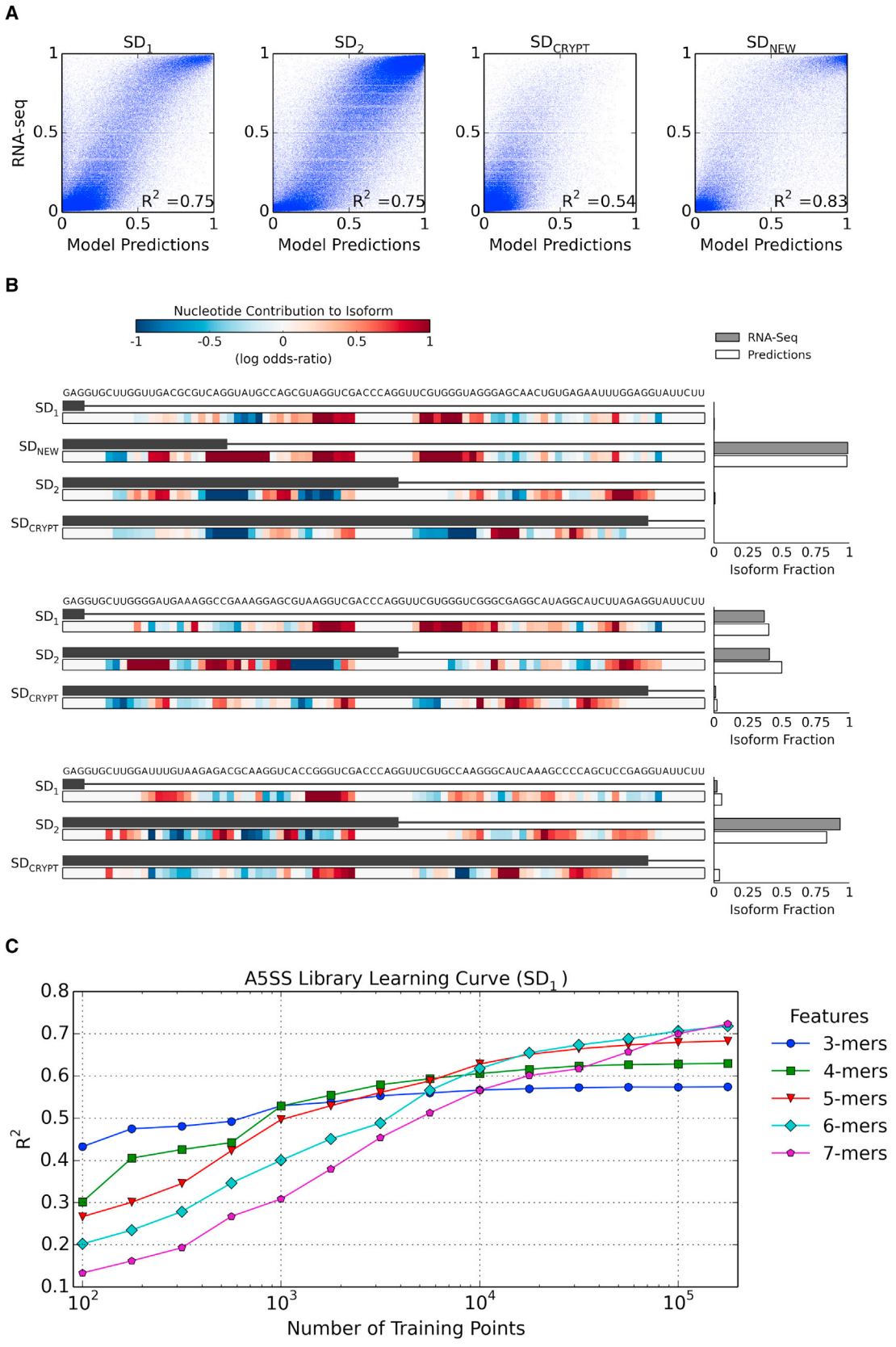
See also Figure S3.

captured by the sum of the 4-mer's individual effect sizes (R^2 = 0.913; Figure 4B). We did the same analysis for 4-mers located in the second degenerate region of the 3′ library. Here, the linear model fit the experimental data even better (R^2 = 0.954; Figure 4C). Thus, while specific instances of cooperative sequence interactions have been well documented (Huelga et al., 2012; Oberstrass et al., 2005), our results suggest the majority of motifs primarily exert their influence on exon inclusion independently of the surrounding motifs.

Predicting Isoform Ratios in Alternative Splicing from Sequence Information

We then turned to the task of learning a model of alternative splicing to predict isoform levels from sequence information. Because combinatorial regulation of alternative splicing was accurately captured by an additive model, we postulated that an additive model with short sequences as input features would perform well for prediction. Using both the 5′ and 3′ libraries, we trained a joint model of alternative exon definition in which a score is learned for each of the 4,096 possible 6-mers (Figure S3A). The scores learned here are similar to the previously calculated effect sizes, but rather than measuring the effects of a single 6-mer one at a time, we learned all the scores together through regression. Given the large number of new splice donors appearing within the 5′ library, we also chose to train a model of the splice donor site itself (Figure S3B). When we tested the splice donor model using cross validation, we found it accurately predicted the fraction of reads mapping to the three original splice donors, accounting for up to 75% of observed isoform variability (R^2: SD_1 = 0.75, SD_2 = 0.75, SD_{CRYPT} = 0.54; Figure 5A). It also proved accurate in predicting the position and fraction of reads mapping to newly created splice donor sites within the degenerate regions (R^2: 0.83; Figures 5A and 5B).

A fundamental advantage of testing synthetic sequences is the ability to learn from larger datasets than were previously available. As an attempt to quantify this advantage, we calculated learning curves on a simple model predicting usage of SD_1 in the alternative 5′ library. We split our data into training and test sets (90%/10% split) and trained models using subsets of the training data (between 100 to 177,827 training points). We also trained separate models using 3-mers, 4-mers, 5-mers, 6-mers, or 7-mers. With limited data (1,000 or fewer training points), the simplest model (3-mers) made the most accurate predictions, while the 7-mer model made the least accurate predictions, with the other models ordering between (Figure 5C). However, with the largest training subset (177,827 points), the results were reversed with the 7-mer model achieving the highest accuracy. Based on the slopes of



(legend on next page)

the learning curves, the 3-mer to 5-mer models would not benefit significantly from more data points (> 177,827), but the 6-mer, and especially the 7-mer, models seem likely to achieve significantly higher prediction accuracy with larger training sets. These results highlight the intuitive point that richer feature sets can improve predictions accuracy, but require more data to properly train.

Predicting the Effects of Human Genomic SNPs on Alternative Isoform Ratios

Next, we asked whether we could apply our model (HAL [hexamer additive linear])—developed entirely in the context of synthetic mini-genes—to predict changes in alternative splice donor usage caused by common polymorphisms in human genomes. As a first test case, we focused on 5′ alternative splicing. Combining DNA and RNA sequencing data, respectively, from the 1000 Genomes Project (Abecasis et al., 2012) and GEUVADIS consortium (Lappalainen et al., 2013), we calculated the percent of splicing at the downstream alternative splice donor (percent spliced in [PSI]) of wild-type genotypes for 8,546 5′ alternative splicing events using the MISO software package (Katz et al., 2010). We separately calculated mean isoform levels for genotypes heterozygous or homozygous for a single SNP in the region between the two competing splice donors or within the splice donors themselves (Table S1).

We began by investigating whether the model of the actual 9-nt splice donor sequence—again learned completely from our synthetic mini-genes—could accurately predict the effects of SNPs occurring within splice donor sequences. We also compared our prediction accuracy to a leading splice donor prediction tool trained directly from splice donor usage in the human genome (MaxEnt) (Yeo and Burge, 2004). Among heterozygous SNPs in alternative splice donors occurring in multiple individuals, we found that 93 of 199 SNPs altered PSI by >5% (Figures 6A and 6B). Within this set, HAL predicted the direction of change with 87.1% accuracy (81/93; binomial $p = 9.83 \times 10^{-14}$), while MaxEnt predicted the direction of change with 81.7% accuracy (76/93; binomial $p = 4.45 \times 10^{-10}$). Among the 35 homozygous SNPs in splice donors that alter PSI by >5%, our model predicted every SNP correctly, while MaxEnt made two mistakes (HAL: 35/35, binomial $p = 5.82 \times 10^{-14}$; MaxEnt: 33/35, binomial $p = 3.67 \times 10^{-10}$). For the set of SNPs within splice donors, our model explained 59.3% of the observed heterozygous effects ($R^2 = 0.593$, $p = 6.38 \times 10^{-8}$) and 67.7% of the observed homozygous effects ($R^2 = 0.677$, $p = 4.65 \times 10^{-24}$). This is a substantial improvement over MaxEnt, which accounted for 39.8% of the observed heterozygous effects ($R^2 = 0.398$, $p = 1.22 \times 10^{-11}$) and 41.1% of the observed homozygous effects ($R^2 = 0.411$, $p = 3.3 \times 10^{-5}$). Even when we extended our analysis to all SNPs (including those with less than 5% change in PSI), we found HAL substantially outperformed MaxEnt (HAL: $R^2 = 0.48$; MaxEnt: $R^2 = 0.22$; Figure S4A).

We then applied the model to predict the effects of human genomic SNPs in the alternative exon region between, but not overlapping, splice donors. Because most SNPs not occurring in actual splice sites are likely to only have modest effects, we restricted our analysis to SNPs with at least ten homozygous wild-type or ten heterozygous samples expressing the relevant mRNA. Moreover, we focused on SNPs that resulted in a change in the PSI of at least 5% to minimize the impact of measurement noise on the validation dataset; 43/344 heterozygous and 20/131 homozygous SNPs altered the PSI by >5% (Figure 6C). HAL correctly predicted the direction of change for 37/43 heterozygous and 17/20 homozygous SNPs (p: heterozygous = 1.63×10^{-6}, homozygous = 2.58×10^{-3}, combined = 6.11×10^{-9}). Furthermore, our model explained around half of the total observed effects of these SNPs (heterozygous: $R^2 = 0.570$, $p = 9.23 \times 10^{-9}$; homozygous: $R^2 = 0.442$, $p = 1.39 \times 10^{-3}$). Thus, our model not only outperformed the state of the art splice donor algorithm (MaxEnt) at predicting the effects of SNPs within splice donors but also successfully predicted the effects of SNPs within the alternative exon region, which to our knowledge, no other tool can do.

Predicting Alternative 5′ Isoform Levels from Sequence Information

To further assess the accuracy of our splice donor model, we predicted the isoform ratios in 6,152 alternative 5′ splicing events expressed in lymphoblastoid cell lines and compared our results to four other splice donor prediction algorithms. Our splice donor model substantially outperformed all of the other algorithms (Figure S5; Table S2). Interestingly, all of the models (including ours) performed better on events with shorter alternative exon regions (i.e., the region between splice donors). In these events, there is less space for regulation between the splice donors, possibly simplifying the prediction task.

Predicting the Effects of Variants on Exon Skipping in Mendelian Diseases

The most common form of alternative splicing is neither alternative 5′ or 3′ splicing, but exon skipping. Exon skipping is a highly regulated form of alternative splicing in human cells, and misregulation of cassette exon splicing can cause disease (Garcia-Blanco et al., 2004) and cancer (Kim et al., 2008). Given the relatively more complex structure of skipped exons, it might on first sight seem unlikely that a model trained only on 5′and 3′ alternative splicing should be able to predict levels of exon inclusion. However, we hypothesized that the similarity between

Figure 5. A Model Accurately Predicts Alternative 5′ Splicing and the Location of New Splice Donors

(A) For each splice donor (SD_1, SD_2, SD_{CRYPT}), model predictions are plotted against the observed splice site usage fraction. Each point represents a single test plasmid. The results are also plotted for all new splice sites (SD_{NEW}).

(B) The prediction results for three different mini-genes are shown with the associated nucleotide scores for each isoform. Each nucleotide score is calculated by averaging the model weights of all 6-mer overlapping the nucleotide. In the first example mini-gene, HAL predicts the usage and position of a new splice donor, which is confirmed by RNA-seq.

(C) A learning curve was generated for different models that predict the fraction of splicing at SD_1. The simplest model (3-mer features) performed the best with small training sets (<1,000 data points), but with more data points, richer feature sets offer better performance.

See also Figure S5.

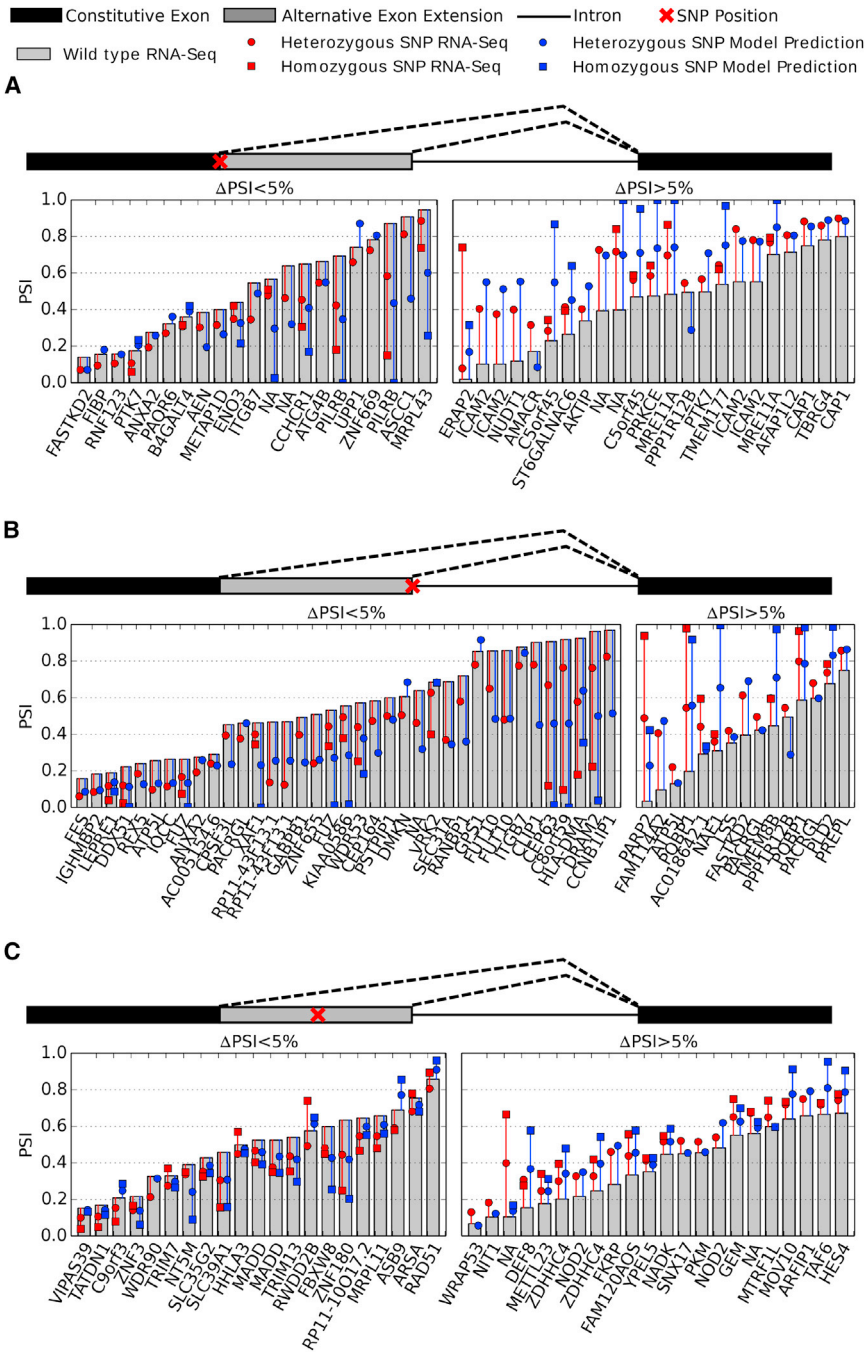


Figure 6. Splicing Model Identifies the Functional Effect of SNPs on Alternative Splicing

(A–C) Model predictions are plotted with the PSI measured from RNA-seq for SNPs occurring in the upstream splice donor (A), the downstream splice donor (B), and between the competing splice donors (C) that alter the measured PSI by greater than 5%. The observed PSI from RNA-seq for the wild-type genotype (gray bar) and genotypes containing the SNP (red) are plotted together with the model prediction (blue). The model accurately predicts the direction of change of the heterozygous SNPs in splice donors with 87.1% accuracy (81/93; binomial $p = 9.83 \times 10^{-14}$) and the heterozygous SNPs between splice donors with 86.0% accuracy (37/43; binomial $p = 8.18 \times 10^{-7}$). See also Figure S4 and Tables S1 and S2.

the sequence determinants of alternative exons in alternative 5′ and 3′ splicing might extend to exon skipping as well. If this were the case, we would expect our model to accurately predict the effects of exonic sequence variants on skipped exon-inclusion levels, even though it was never trained directly on any exon skipping data. We tested this hypothesis in the context of mutations in several distinct genes that are known to cause Mendelian disease by promoting exon skipping (Figure 7A; Table S3).

First, we compared model predictions to experimental data for the *SMN1* and *SMN2* genes, whose misregulation can lead to spinal muscular atrophy. Our model correctly predicted increased or decreased exon 7 inclusion in 205/229 (89.5%; Figure 7D) variants with experimental data. In Figure 7B, we compare predictions (increased or decreased exon inclusion) to experimental data. To make the plot more readable, we only included a single SNP at each position. Our model accurately predicts increased/decreased exon inclusion for 20/22 of the plotted SNPs. On just the variants with quantitative data (n = 131), our model explained 65% of the observed variance (R^2 = 0.65; Figure 7E). The *SMN1/2* variants that we tested included SNPs, indels, and combinations of up to 30 nt changes.

We then tested our model on variants in *CFTR*, whose misregulation can lead to cystic fibrosis. Our model correctly predicted increased/decreased exon 12 inclusion in 19/22 variants (Figure 7D). When we only looked at the SNP with the largest effect at each position, our model accurately predicted increased/decreased exon inclusion for 11/12 SNPs (Figure 7C). Among all the *CFTR* variants, our model explained 60% of the observed variance (Figure 7E; R^2 = 0.60).

Next, we tested our model predictions on variants in exon 7 of the *BRCA2* gene, a tumor suppressor responsible for DNA damage repair. Mutations in *BRCA2* affecting the ability of the protein to repair DNA lead to such an increased risk of ovarian and breast cancer that patients with these mutations may choose to have prophylactic surgery. However, the effect of many variants on alternative splicing and hence protein function remain unknown, forcing patients and doctors to make clinical decisions with limited information. The ability to identify deleterious variants computationally can provide valuable information to

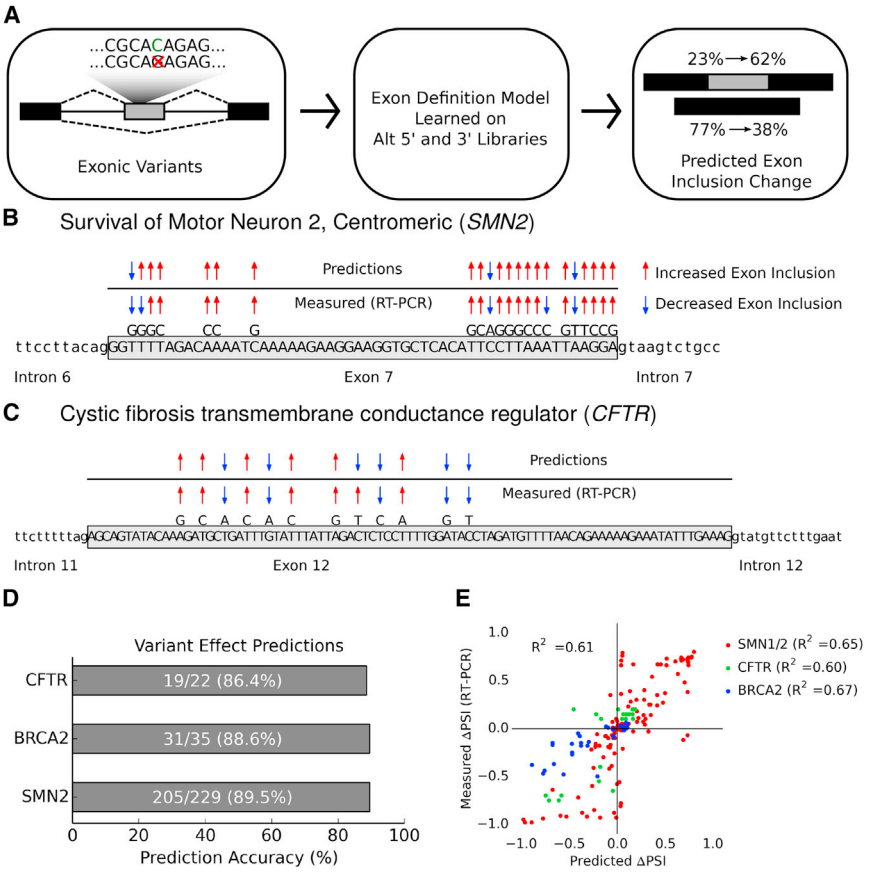


Figure 7. Predicting the Effects of Exonic Variants on Exon Skipping

(A) The inputs to the splicing model can include SNPs, indels, or complex variants within the alternative exon. The splicing model then predicts the exon inclusion levels with the variant present.

(B) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 7 of SMN2. For positions with data for multiple SNPs, the SNP with the largest measured change in PSI was plotted. The model accurately predicted the directional change in PSI (increased exon inclusion/exclusion) for 20/22 SNPs plotted.

(C) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 12 of CFTR. The model accurately predicted the directional change in PSI for 11/12 SNPs plotted.

(D) The prediction accuracy for variants in SMN2, CFTR, and BRCA2 ranged from 86% to 90%.

(E) The change in PSI is plotted for every variant with RT-PCR data. The model explains over 60% of the effects of SNPs for variants each gene tested (SMN1/2, CFTR, and BRCA2).

See also Figure S6 and Table S3.

patients with these variants of unknown significance. Our model correctly predicted increased/decreased exon 7 inclusion for 31/35 variants that experimentally altered inclusion levels (Figure 7D). The model correctly predicted 19/22 of the SNPs with the largest effect at each position within the exon (Figure S6B). Among all the *BRCA2* variants, our model explained 67% of the observed variance (R^2 = 0.67; Figure 7E).

We then compared our results to SPANR (Xiong et al., 2014)—the current state of the art in predicting the effects of SNPs on exon skipping. SPANR consists of a Bayesian deep learning algorithm trained on exon skipping events in the human genome with 1,393 carefully hand-selected features. As of this paper, SPANR only supports predictions of SNPs, so we were not able to compare our predictions on more complex variants. However, for SNPS in *SMN1/2*, *CFTR*, and *BRCA2*, we found that HAL accounted for three times more of the observed effects than SPANR (HAL: R^2 = 0.51; SPANR: R^2 = 0.17; Figure S6A). We made HAL publicly available at http://splicing.cs.washington.edu. All of the code to reproduce this study is publicly available at https://github.com/Alex-Rosenberg/cell-2015.

DISCUSSION

We present a framework based on massively parallel analysis of synthetic sequences to dramatically improve our understanding of alternative splicing and the ability to predict the impact of natural human genetic variation. Our model accurately predicts the effects of sequence variants on alternative 5′ splicing that occur both within the alternative exon and in the competing splice donors. Even more importantly, our model learned regulatory rules about alternative splicing that generalized to exon skipping—a completely different form of alternative splicing than those on which the model was trained.

Our results suggest that a common regulatory mechanism is shared between all major forms of alternative splicing. Additional evidence for such a common mode of regulation comes from previous smaller-scale studies of ESEs or ESSs that have shown similar effects across different forms of alternative splicing (Wang et al., 2006, 2012). It is unlikely that this shared form of regulation occurs during splice site recognition; any exonic splice regulatory element that alters splice donor or splice acceptor recognition should have different effects in alternative 5′ and 3′ splicing events. It is more likely that alternative exon inclusion is modulated during exon definition, that is the pairing of splice site across exons, which often precedes the eventual pairing of splice donors and acceptors across introns (Robberson et al., 1990).

Furthermore, our data also suggest that the exon-defining interactions between the upstream splice acceptor and downstream splice donor are regulated additively. In both alternative 5′ and 3′ splicing, we found the joint effect size of multiple 4-mer to be highly correlated with the sum of the individual 4-mer effects. This result may indicate that each sequence motif can contribute additively to stabilizing the splice acceptor-splice donor interaction, likely through the *trans*-factors that bind these sites. However, the true

mechanistic basis for this additivity will require further investigation. Although, there is evidence supporting specific examples of functional interactions between *cis*-splicing regulatory elements (Oberstrass et al., 2005), our results indicate that these examples are likely uncommon.

A potential limitation of our approach is that mRNAs are transcribed from plasmids rather than directly from the genome, especially considering evidence suggesting that chromatin can influence alternative splicing (Luco et al., 2010). However, advances in high-throughput genome editing may make it possible to perturb the genome in a massively parallel fashion, which will enable extensions of our approach to probe the effects of chromatin on alternative splicing. In fact, recent work demonstrated that small-scale genomic libraries could be created through insertion of degenerate sequences directly into an alternatively spliced gene locus (Findlay et al., 2014). Moreover, our current work focused on mini-genes with short alternative exons, and more work will be necessary to understand to which extent our results generalize to other gene architectures. However, human exons are typically short (an average 147 bp for internal exons) (IHGSC et al., 2001), and, moreover, analysis of sequence conservation suggests that most sequence determinants of alternative splicing can be found within a few hundred nucleotides of intron-exon junctions. It is important to emphasize that our approach uncovers only *cis*-regulatory rules. Complementary experiments that connect this *cis*-grammar to a repertoire of *trans*-acting splice factor proteins are necessary to fully understand the mechanisms underlying the regulation of alternative splicing.

We have demonstrated that learning the sequence determinants of gene regulation from large libraries of synthetic sequences can be used as a complementary approach to learning directly from the human genome. We assayed over two million alternatively spliced constructs, nearly two orders of magnitude more events than the 38,000 that are present in the human genome (Wang et al., 2008), containing over 100 Mb of synthetic sequence. Our improved understanding of alternative splicing and performance in predicting the effects of genetic variants is not a result of more sophisticated machine learning algorithms but simply the result of learning from a larger and more reliable dataset. We anticipate that this general approach will be useful for advancing our biological understanding of diverse forms of gene regulation, such as transcription, translation, and polyadenylation.

EXPERIMENTAL PROCEDURES

Cloning of Degenerate Libraries

The libraries were assembled with PCR and standard Gibson assembly (Gibson et al., 2009) using degenerate oligonucleotides (IDTDNA). First Citrine was split into two exons, and the first exon of the Citrine gene was altered to remove any potential splice donors, without altering the amino acid sequence. The introns with degenerate sequences were inserted between the two exons of Citrine. The barcode sequence was inserted into the 3′ UTR of Citrine.

Cell Culture and Transfection

HEK293 cells were cultured in in DMEM (Cellgro) plus 10% FBS and L-glutamine/penicillin/streptomycin on coated plates. Plates were coated for 24 hr with 8 ml of 100× diluted extracellular matrix gel (Sigma-Aldrich) before HEK293 cells were added to the plates. For transfection of a complex pool of plasmids, 1.2 million cells were seeded in a 10-cm dish 24 hr before transfection. We mixed 10 μg of the plasmid library in 1 ml of Opti-MEM Reduced Serum Medium (Life Technologies) with 30 μl of Lipofectamine LTX and 10 μl of Plus Reagent (Life Technologies), before transfecting into the 10-cm dish. The DMEM was replaced 5 hr after transfection.

Isolation of RNA and Generation of cDNA

Total RNA was extracted using RNeasy (QIAGEN) kits 24 hr after transfection. The optional on column DNaseI digest was performed with the RNase-Free DNase Set (QIAGEN). Total RNA quality and purity was tested by measuring the A260/A280 ratio on a NanoDrop 1000 Spectrophotometer and, in some cases, by measuring the ratio of the 18S and 28S rRNA bands on a native 1% agarose gel. mRNA was separated from 35–48 μg total RNA using polyA Spin mRNA Isolation Kits (New England Biolabs). Isolated mRNA was again digested by DNaseI for 30 min using the Turbo DNA-free Kit (Ambion). cDNA was then synthesized from 109–374 ng mRNA using MultiScribe Reverse Transcriptase (Ambion) and Oligo d(T)16 primers (Ambion). cDNA synthesis was performed by holding reactions at 25°C for 10 min, 42°C for 110 min, and 85°C for 5 min. The quality of cDNA and presence of DNA contamination were checked through qPCR: Citrine, mCherry, and TBP were compared using cDNA, no reverse transcription controls (NRTC), and a no template control (NTC). The results indicated that there was no plasmid or genomic DNA carryover into the cDNA reactions.

Generation of Illumina Flow Cell Compatible PCR Products from RNA and DNA Library

The resultant cDNA was then amplified by PCR to generate products compatible with the Illumina HiSeq2000 Flow Cell. PCR reactions were performed in 100 μl with 2x Phusion HF Master Mix (New England Biolabs), 50 pmol forward primer, and 50 pmol reverse primer with sample specific barcodes and 20% of each cDNA reaction. Cycling was done on a BioRad T100 Thermal Cycler with the following protocol: 98°C for 5 min, then seven cycles of 98°C for 10 s, 67.5°C for 15 s, 72°C for 30 s, and a final extension step at 72°C for 5 min. The necessary number of cycles was determined for each sample by first running qPCR reactions with EvaGreen in a Biorad CFX and determining when fluorescence began to plateau. Following PCR, 10% of the products were run on a 2% agarose gel to determine if the expected bands were present. The remainder of the PCR products was purified using the QIAquick PCR Purification Kit (QIAGEN) and eluted into 30 μl of EB. Concentrations, as well as A260/280 and A260/230 ratios, were measured on a NanoDrop 1000 Spectrophotometer.

Illumina-compatible PCR products were also generated from the DNA plasmid library with the same protocol as above, except the cDNA template was replaced with 10 ng of plasmid library DNA and the PCR reaction was performed with 20 cycles.

Sequencing Plasmid Library and RT-PCR Products

Both the RT-PCR products and plasmid library PCR products were sequenced on either an Illumina HiSeq2000 or Illumina MiSeq with paired end reads. The forward read crossed the post-splicing exon-exon junction and the reverse read covered the 3′ UTR barcode. A 6-nt index read was used to sequence the sample barcode to determine if the read came from a DNA library or a cDNA library.

Associating Degenerate Intronic Regions with 3′ UTR Barcode Tags

Using the sequencing results of the DNA plasmid library, we first counted the number of reads for every observed barcode and calculated an average Phred quality score for each position. We discarded any barcode tags with less than two reads or less than an average Phred score of 20 at any position. We then mapped each remaining tag to the associated degenerate sequence with the most reads. If each degenerate sequence had a single read, we chose the sequence with the highest minimum Phred score.

Measuring Isoform Fractions from Sequencing Results

For every read on an RT-PCR product, we recorded the splicing position (or lack of splicing) by aligning the read to the unspliced plasmid. Using the associated barcode read, we were then able to tally the number of reads splicing

at each position for every plasmid in our library. With respect to the alternative 5′ library, only reads that mapped to a splice donor with GT or GC in the +1 to +2 intronic positions were counted.

ACCESSION NUMBERS

The accession number for the raw and processed sequencing data reported in this paper is GEO: GSE74070.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.054.

AUTHOR CONTRIBUTIONS

A.B.R. designed and performed experiments, analyzed data, built and tested the splicing model, wrote the manuscript, and developed the web tool; R.P.P. designed experiments; and J.S. and G.S. designed the experiments and wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by a National Science Foundation (NSF) Career Award (0954566) and a Burroughs Wellcome Career Award at the Scientific Interface (to G.S.).

Received: July 14, 2015
Revised: August 28, 2015
Accepted: September 21, 2015
Published: October 22, 2015

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.

Culler, S.J., Hoff, K.G., Voelker, R.B., Berglund, J.A., and Smolke, C.D. (2010). Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. Nucleic Acids Res. *38*, 5152–5165.

Doma, M.K., and Parker, R. (2006). Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature *440*, 561–564.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Garcia-Blanco, M.A., Baraniak, A.P., and Lasda, E.L. (2004). Alternative splicing in disease and therapy. Nat. Biotechnol. *22*, 535–546.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. Mol. Cell *22*, 769–781.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. *1*, 167–178.

International Human Genome Sequencing Consortium, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Kim, E., Goren, A., and Ast, G. (2008). Insights into the connection between cancer and alternative splicing. Trends Genet. *24*, 7–10.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Le, Q.V., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., and Ng, A.Y. (2012). Building high-level features using large scale unsupervised learning. In Proceedings of International Conference in Machine Learning.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA *100*, 189–192.

Listerman, I., Sapra, A.K., and Neugebauer, K.M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat. Struct. Mol. Biol. *13*, 815–822.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Martinez-Contreras, R., Fisette, J.-F., Nasim, F.U., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol. *4*, e21.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Noderer, W.L., Ross, J.F., Aparna, B., Alexander, J.D.A., Jiajing, Z., Paul, A.K., and Clifford, L.W. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol. Syst. Biol. *10*, 1–14.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., and Allain, F.H. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science *309*, 2054–2057.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell Rep. *7*, 281–292.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Shoemaker, C.J., Eyler, D.E., and Green, R. (2010). Dom34:Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. Science *330*, 369–372.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580–586.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wang, Z., Xiao, X., Van Nostrand, E., and Burge, C.B. (2006). General and specific functions of exonic splicing silencers in splicing control. Mol. Cell *23*, 61–70.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol. *19*, 1044–1052.

Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat. Struct. Mol. Biol. *20*, 36–45.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. USA *110*, 11952–11957.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2014). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, no. 6218.

Yeo, G., and Burge, C. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yu, Y., Maroney, P.A., Denker, J.A., Zhang, X.H., Dybkov, O., Lührmann, R., Jankowsky, E., Chasin, L.A., and Nilsen, T.W. (2008). Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. Cell *135*, 1224–1236.

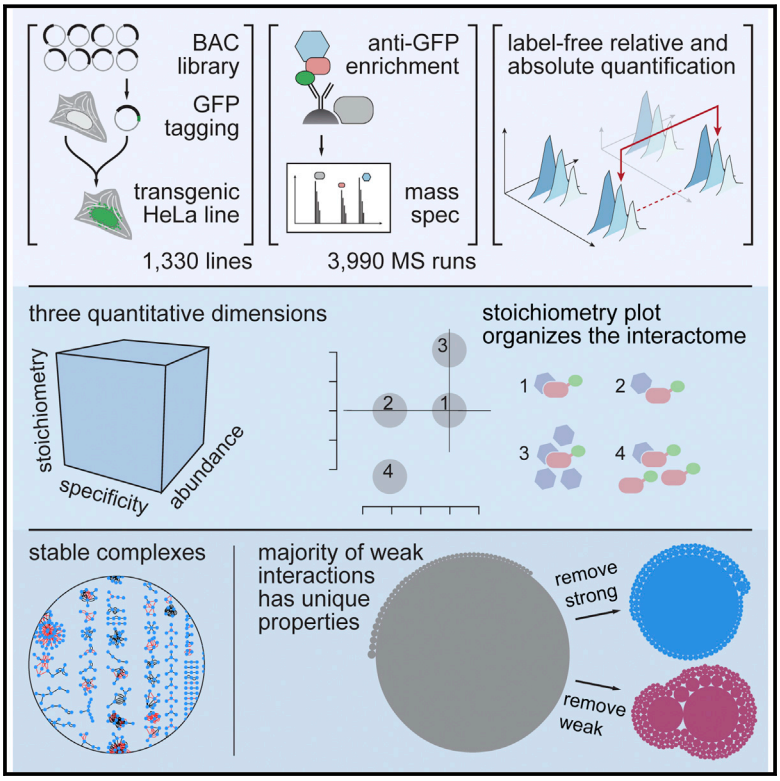
Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18*, 1241–1250.

Article

Cell

A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances

Graphical Abstract



Highlights

- Human interactome dataset connecting 5,400 proteins with 28,500 interactions
- Three quantitative dimensions measure specificities, stoichiometries, and abundances
- Stable complexes are rare but stand out by a signature of balanced stoichiometries
- Weak interactions dominate the network and have critical topological properties

Authors

Marco Y. Hein, Nina C. Hubner, Ina Poser, ..., Frank Buchholz, Anthony A. Hyman, Matthias Mann

Correspondence

hyman@mpi-cbg.de (A.A.H.), mmann@biochem.mpg.de (M.M.)

In Brief

Weak interactions shape the cellular protein interaction network as determined from proteomic measures of cellular interaction specificities, the strength of those interactions, and the cellular copy numbers of the proteins involved.

Hein et al., 2015, Cell *163*, 712–723
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.053

CellPress

Cell

Article

A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances

Marco Y. Hein,[1,6,8] Nina C. Hubner,[1,6,9] Ina Poser,[2] Jürgen Cox,[1] Nagarjuna Nagaraj,[1] Yusuke Toyoda,[2,10] Igor A. Gak,[3] Ina Weisswange,[4,5] Jörg Mansfeld,[3] Frank Buchholz,[2,4] Anthony A. Hyman,[2,7,*] and Matthias Mann[1,7,*]

[1]Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany
[3]Cell Cycle, Biotechnology Center, TU Dresden, 01307 Dresden, Germany
[4]Medical Systems Biology, UCC, Medical Faculty Carl Gustav Carus, TU Dresden, 01307 Dresden, Germany
[5]Eupheria Biotech GmbH, 01307 Dresden, Germany
[6]Co-first author
[7]Co-senior author
[8]Present address: Howard Hughes Medical Institute, University of California San Francisco, San Francisco, CA 94143, USA
[9]Present address: Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, Nijmegen 6525 GA, the Netherlands
[10]Present address: Institute of Life Science, Kurume University, Kurume, Fukuoka 839-0864, Japan
*Correspondence: hyman@mpi-cbg.de (A.A.H.), mmann@biochem.mpg.de (M.M.)
http://dx.doi.org/10.1016/j.cell.2015.09.053

SUMMARY

The organization of a cell emerges from the interactions in protein networks. The interactome is critically dependent on the strengths of interactions and the cellular abundances of the connected proteins, both of which span orders of magnitude. However, these aspects have not yet been analyzed globally. Here, we have generated a library of HeLa cell lines expressing 1,125 GFP-tagged proteins under near-endogenous control, which we used as input for a next-generation interaction survey. Using quantitative proteomics, we detect specific interactions, estimate interaction stoichiometries, and measure cellular abundances of interacting proteins. These three quantitative dimensions reveal that the protein network is dominated by weak, substoichiometric interactions that play a pivotal role in defining network topology. The minority of stable complexes can be identified by their unique stoichiometry signature. This study provides a rich interaction dataset connecting thousands of proteins and introduces a framework for quantitative network analysis.

INTRODUCTION

Proteins are central protagonists of life at the molecular level. They interact for structural, regulatory, and catalytic purposes, forming macromolecular structures as well as stable or transient multi-protein complexes. Accordingly, protein interactions vary greatly in their biophysical properties, while protein abundances range from a few to millions of copies per cell. The interactome is therefore the product of two factors: binary affinities between protein interfaces (Rual et al., 2005; Stelzl et al., 2005; Rolland et al., 2014) and the cellular proteome, which itself is characterized by subcellular localization, post-translational modifications and protein concentrations (Hein et al., 2013; Mann et al., 2013).

Mapping the protein interactome landscape has been a long-standing goal of modern biology and a variety of methods have been developed to this end (Seebacher and Gavin, 2011). Affinity purification followed by mass spectrometry (AP-MS) can in principle determine the members of protein complexes in their cellular context in an unbiased manner (Gingras et al., 2007) and has enabled large-scale protein interaction studies of several model organisms, including human cells (Ewing et al., 2007; Malovannaya et al., 2011). Nanoscale liquid chromatography (LC) coupled to sensitive and fast mass spectrometers has boosted interaction proteomics technology in recent years, increasing coverage and minimizing false negative rates. It has also enabled a paradigm shift from identification to quantification of interacting proteins (Bantscheff et al., 2012). Quantitative approaches permit the use of mild immunoprecipitation (IP) protocols and allow specific binders to stand out by their quantitative signature even from very large backgrounds of unspecific proteins (Mellacheruvu et al., 2013; Keilhauer et al., 2015). Additionally, MS-based proteomics is now able to characterize entire cellular proteomes with increasingly complete coverage (Beck et al., 2011; Mann et al., 2013), providing abundances and copy-number estimates of the expressed proteins. This should now allow studying the quantitative interactome as a function of the underlying proteome. To generate model systems that closely recapitulate in vivo conditions, we have previously developed bacterial artificial chromosome (BAC) transgeneomics: GFP-tagged proteins are expressed in mammalian cell lines from BAC transgenes with near-endogenous expression patterns from human or orthologous mouse loci (Poser et al., 2008). GFP-based tags are dual-purpose in that they can be used for both imaging and as affinity handle. Combining these

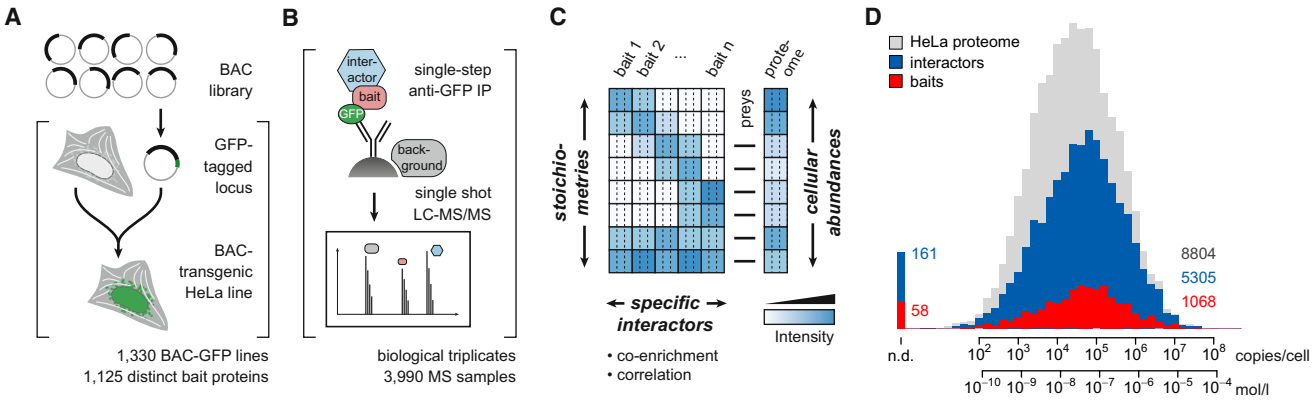


Figure 1. Quantitative BAC-GFP Interactomics

(A) BAC recombineering workflow for generating transgenic HeLa lines.

(B) Single-step affinity-purification, single-run liquid chromatography-tandem mass spectrometry (LC-MS/MS) workflow.

(C) Schematic protein quantification matrices in interactome and proteome samples with three dimensions of quantification.

(D) Proteome coverage and abundance distribution of the bait proteins and their interactors.

See also Tables S1, S2, and S3.

cell lines with the quantitative proteomics workflow resulted in a versatile and highly specific method that we termed quantitative BAC-GFP interactomics (QUBIC) (Hubner et al., 2010).

Here, we applied QUBIC in a proteome-wide manner, using 1,125 bait proteins to assemble a large-scale map of the human interactome. We characterize individual interactions in three quantitative dimensions that address statistical significance, interaction stoichiometry, and cellular abundances of interactors. This concept provides a unique perspective on the interactome, enabling the discovery and characterization of stable and transient protein complexes, guiding their functional interpretation and shedding light on the topological architecture of the entire network.

RESULTS

Quantitative BAC-GFP Interactomics

Collections of strains or cell lines expressing tagged proteins are indispensable tools for many systems biology approaches (Huh et al., 2003). Expressing GFP-tagged proteins from engineered BAC transgenes maintains the endogenous promoters, intron-exon-structures and regulatory elements, ensuring near-endogenous expression levels and patterns (Poser et al., 2008) (Figure S1A). We have previously used this system to study chromosome segregation and the function of motor proteins (Hutchins et al., 2010; Maliga et al., 2013). To map the protein interactome globally, we generated a resource of 1,330 stable BAC-GFP HeLa cell lines (Figure 1A; Table S1). Mouse BACs are excellent surrogates for their human orthologs and offer additional options, such as resistance to RNAi against their endogenous counterparts, streamlining functional studies of the tagged proteins (Kittler et al., 2005). In 615 cell lines, we used mouse BACs with a median sequence identity of 94% with their respective orthologs (Figure S1B). Overall, our collection encompasses 1,125 distinct bait proteins across all protein classes (Figures S1C–S1E), some present as C- and N-terminally tagged versions, or as mouse and human sequences (not counted as distinct).

We performed QUBIC in three biological replicate experiments, resulting in 3,990 LC-MS runs recorded on an Orbitrap mass spectrometer, taking about a year of net measuring time (Figure 1B). To define specific interactors, we employed MaxLFQ, the label-free quantification (LFQ) module of the MaxQuant software (Cox and Mann, 2008; Cox et al., 2014). Bait proteins and their interactors are characterized by quantitative co-enrichment compared to their intensity profiles across many samples (Figures 1B and 1C), and we used generic statistical testing to determine significantly enriched cases (Keilhauer et al., 2015). To set thresholds for accepting a given candidate as an interactor, we developed an entirely data-driven, false discovery rate (FDR)-controlled approach that harnesses the absence of "negative" interactions and the concomitant asymmetry of the outlier population (Figures S1F and S1G). This approach does not rely on reference datasets or prior knowledge for training, but nonetheless validates favorably against gold standards (Figures S1I–S1K).

In addition to local co-enrichment, we found the intensity profiles of interacting proteins to be closely correlated globally (Supplemental Experimental Procedures). Profile correlations alone can indicate protein interactions when proteome samples are subjected to extensive native fractionation (Havugimana et al., 2012; Kristensen et al., 2012). Here, we use them as additional classifiers (Keilhauer et al., 2015) and the combination of enrichment FDRs and profile correlation coefficients defines the confidence class of each interaction (Figure S1H).

Overall, using the information in this first dimension of proteomic quantification, our analysis resulted in 28,504 unique and statistically significant interactions involving 5,462 distinct proteins (Table S2).

Interaction Stoichiometries and Protein Abundances

A second dimension of quantification can in principle be applied to determine the stoichiometries of proteins within complexes.

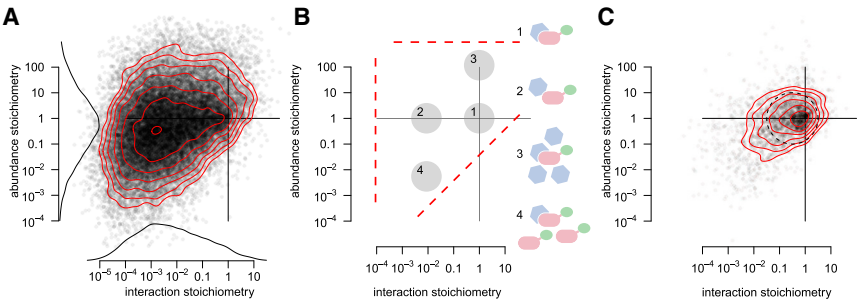


Figure 2. The Stoichiometry Plot

(A) Overlay of all interaction and abundance stoichiometry data for all interactions.

(B) The characteristic triangular shape is a consequence of the dynamic range limits in the interactome (left border), in the proteome (top border) and the stoichiometry limit imposed by the relative cellular protein abundances (diagonal). Schematic interaction scenarios: (1) equal cellular abundance, stable interaction; (2) equal cellular abundance, weak interaction; (3) stable interaction with greater cellular abundance of the prey; and (4) reciprocal case: quantitative recovery of a stably bound, less abundant prey.

(C) Stoichiometry plot of interactions between proteins annotated as CORUM complex members. The area of highest density can be approximated by a circle containing 58% of CORUM interactions.

See also Data S1.

These can be computationally extracted from label-free affinity purification data with accuracies reaching those of methods using isotopically labeled reference standards (Wepf et al., 2009; Smits et al., 2012). If a protein complex contained one copy of each subunit, one might expect them to be retrieved in equimolar amounts after immunoprecipitation (IP). However, in practice, measured stoichiometries between preys and baits span orders of magnitude (Collins et al., 2013; Hauri et al., 2013). This is because the observed stoichiometries depend on more than the initial composition of the individual complexes in the cell. For instance, limited kinetic and thermodynamic stability can result in substoichiometric recovery. Proteins may also reside in different alternative molecular assemblies with fractions of their total cellular pools. Hence, we hypothesized that globally, interaction stoichiometries might reflect the stability of a given protein-protein interaction and depend on the extent that interactors are engaged with each other. The cellular abundance of an interactor can be limiting for how much is recoverable after immunoprecipitation (IP), setting a lower bound for the interaction stoichiometry. We therefore reasoned that cellular copy numbers would provide a crucial third quantitative dimension.

For each pair of interacting proteins, we first quantified their stoichiometry in the immunoprecipitates using a sophisticated label-free strategy for absolute protein quantification (Supplemental Experimental Procedures). To determine the precision of our method, we systematically compared interaction stoichiometries from experiments where the same bait proteins were tagged on different termini, representing entirely separate experiments (Figures S2A and S2E). Stoichiometries showed high correlation, precision within a factor of three, and no systematic bias for a given terminus. This confirmed that our approach robustly delivers interaction stoichiometries in high throughput; however, these may not always be sufficiently accurate to reliably specify copy numbers of each subunit, even for stable complexes. We repeated the analysis for cases where either the mouse or human ortholog of a protein was used as bait, demonstrating the same level of reliability and no species bias (Figures S2B–S2D). This highlights the extraordinary degree of conservation of protein function in evolutionary time (Kachroo et al., 2015) and suggests that our human-centric dataset is representative not only of the human but other mammalian species.

To add a third dimension of proteomic quantification, we next performed a whole proteome quantification experiment on the parental HeLa cell line that all our BAC-transgenic lines are derived from, to a depth of about 9,000 proteins. To estimate cellular protein abundances, we applied our label-free approach and scaled the values to copies per cell using the "proteomic ruler" concept (Wiśniewski et al., 2014) (Supplemental Experimental Procedures). The proteome dataset provided cellular copy numbers for 5,305 proteins of the interactome dataset, covering 97% of all interactors (Figure 1D). The abundances of interacting proteins closely follow the distribution of bait abundances, covering the entire dynamic range of the proteome. This demonstrates that our BAC-based system recapitulates the in vivo situation, enabling us to probe the interactome as a function of the endogenous cellular proteome.

Quantifying the Interactome in Two Additional Dimensions

Having established a set of specific interactions with the first dimension of quantification, we next combined our second and third dimension of quantification, namely the interaction stoichiometries and relative cellular abundances of interactors. A plot of the stoichiometry landscape for each bait protein is a powerful tool to organize its interactome, because each region reflects a different scenario (Figures 2A and 2B): stable, one-to-one, and fully recovered complexes in which the partners have equal cellular abundance appear around the origin of the plot (case 1 in Figure 2B). Superstoichiometry, the recovery of more prey than bait, is only expected for stable complexes containing more prey than bait copies and indeed we find few of these. If interactions are weak and complexes dissociate partially during IP, or if interactions involve only part of the bait pool, interactors are recovered at substoichiometric levels (case 2), reflecting lower occupancy of interaction interfaces of the bait. A vast predominance of sub- over superstoichiometry confirms our initial hypothesis that stability and occupancy are the main determinants for most interactions.

We observed many cases of stable interactors (~1:1 interaction stoichiometry) that involved a more abundant prey (case 3), such as the interaction of the abundant GTP-binding protein RAN with its guanine-nucleotide releasing factor RCC1 or that of α-tubulin with the NEK9 kinase (see Table S2). The reciprocal

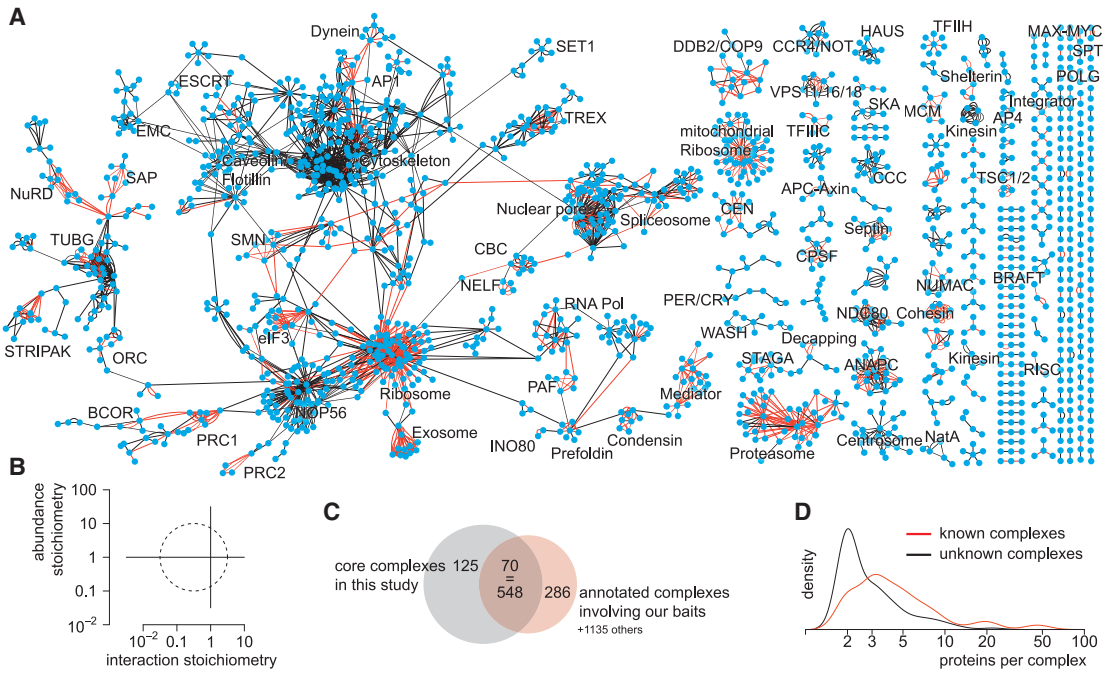


Figure 3. The Core-Complex Network

(A) Sub-networks of interactions matching the CORUM-characteristic core stoichiometry signature. Red edges are known interactions annotated in UniProt or CORUM.

(B) Graphical definition of the core stoichiometry signature. Center: −0.5, 0; radius: 1 ($\log_{10}$ units).

(C) Quantification of the CORUM overlap. A total of 125 isolated networks remain unannotated; 70 networks are annotated with 548 partially redundant CORUM terms; 286 terms assigned to our baits were not shared by any interactor.

(D) Size distribution of annotated versus unannotated networks.

See also Figure S3.

interaction stoichiometry readouts are necessarily smaller than one, because any higher abundant bait can maximally recover the entire pool of its lower abundant prey (case 4). (Note that this would be the default case for overexpressed baits.) We retrieved substoichiometric interactions over an estimated five orders of magnitude; for example, NEK9 was recovered at $6 \times 10^{-6} \times$ the amount of α-tubulin. The proteome-interactome relationship requires that interactors can only be recovered to the extent permitted by their abundance and translates into a diagonal cut-off in the plot, which results in a characteristic triangular shape of the "cloud" of interactions (Figure 2A).

Approximately 10% of our interactions connected members of well-characterized complexes annotated in the CORUM database (Ruepp et al., 2010). They populate a confined area characterized by a signature of balanced stoichiometries (case 1 in Figure 2B). Thus the prototypical case of a stable protein complex as typically described in the literature mostly consists of proteins of equal cellular abundances that are all constitutively bound to each other.

Extrapolating from the signature of known complexes, we reasoned that deduction of similar complexes should be possible solely from the stoichiometry signature of individual baits as opposed to analysis of the entire network (Collins et al., 2007; Hart et al., 2007). We filtered our data for those featuring the core stoichiometry signature (Figure 3B), yielding a larger cluster connecting several molecular assemblies such as major cytoskeletal proteins, the nuclear pore complex and the ribosome as well as 194 isolated putative core complexes (Figure 3A). These recapitulated the majority of CORUM-annotated complexes that involve our bait proteins (Figure 3C). We confirmed the known tendency of large complexes to be well annotated (Havugimana et al., 2012), while smaller assemblies lacked previous description (Figure 3D). The largest of our 125 networks with no database annotation at the time is the recently discovered COMMD/CCDC22/CCDC93 (CCC) complex (Phillips-Krawczak et al., 2015).

The stoichiometry plot offers a unique opportunity for comparing the overlap of our dataset with published data (Figures S3A–S3C). For instance, the intersection with a recent co-fractionation interactome study (Havugimana et al., 2012) closely recapitulated the core-complex signature, with 26% of our core-interactions overlapping with that study. This indicates that the co-fractionation methodology offers an attractive short-cut to finding stable, obligate core complexes. Conversely, the overlap with iRefWeb, a portal of consolidated protein interactions from different sources (Turner et al., 2010), reached much further into the substoichiometric region, beyond stable complexes, but still only covered 16% of our dataset. Finally, the overlap with recent large-scale yeast-two-hybrid data (Rolland et al., 2014) was low (0.4%) and mostly limited to cases characterized by quantitative

prey recovery to the extent permitted by cellular abundance. Moreover, the stoichiometry plot quantitatively confirmed the intuitive notion that high-stoichiometry interactions are easier to detect as they are enriched in the 1% FDR compared to the 5% FDR cohort (Figures S3D and S3E). This is also reflected in the overlap of gene ontology (GO) annotations in pairs of interacting proteins (Figure S3F).

Interactions Explain Phenotypes and Genetic Associations

Our dataset provides an extensive resource that can be mined for new or poorly characterized protein interactions. For instance, among the interactors of SUCO, one other protein, TAPT1, stood out by its core stoichiometry signature, suggesting a novel, stable complex consisting of these two low abundant integral membrane proteins of the ER (Figures 4A, 4B, and S4A–S4C). Mutants of their murine orthologs exhibit severe defects during skeletal development: Truncation of TAPT1 causes transformations in the axial skeleton and perinatal lethality (Howell et al., 2007), whereas loss of the SUN domain-containing ossification factor SUCO (also known as OPT) impairs postnatal bone formation, causing fractures and neonatal death (Sohaskey et al., 2010). The latter study linked the phenotype to impaired rough ER expansion and consequent failure of osteoblasts to secrete collagen required for bone formation. Knockdown of human SUCO increased the cells' resistance against ricin, whose toxicity depends on endocytosis and retrograde trafficking to the ER (Bassik et al., 2013). Similarly, the yeast ortholog of TAPT1, EMP65 (YER140W), is involved in protein folding in the ER and shows buffering genetic and physical interactions with the SUN domain protein SLP1 (YOR154W) (Jonikas et al., 2009; Friederichs et al., 2012). We used our interaction methodology on GFP-tagged strains to confirm this complex (Figures S4D and S4E). Similarly, we validated the reciprocal interaction in the mammalian system using TAPT1 as bait (Figure S4B). Together, our findings establish TAPT1–SUCO as the higher eukaryote ortholog of SLP1-EMP65: a low abundant ER membrane complex that is required for normal skeletal development.

Going beyond stable complexes, we discovered an interaction between the anaphase promoting complex or cyclosome (APC/C) and the uncharacterized protein KIAA1430. The stoichiometry plot indicated that KIAA1430 is of lower cellular abundance and is not an obligate member of the APC/C, as the partners were recovered substoichiometrically at ~1% of the respective baits in reciprocal experiments (Figures 4C and 4D).

To independently test whether KIAA1430 was indeed a transient interactor of the APC/C, we performed a purify-after-mixing (PAM)-SILAC experiment (Wang and Huang, 2008) (Figures 4E and 4F; Supplemental Experimental Procedures). We mixed differentially SILAC-labeled lysates from tagged and control cell lines before the affinity step. Subsequently measured SILAC ratios are indicative of the stability of the interaction, because transient interactors exchange dynamically with unbound counterparts, shifting their ratio toward unity, whereas stable interactors maintain their label ratio. Our results confirmed that only known subunits of the APC/C are stably bound to the core subunit CDC23. Consistently, only some of them were recovered when assayed for binding to KIAA1430 and with ratios indicating a high degree of dynamic exchange. Next, we tested whether KIAA1430 is a substrate of the APC/C by monitoring its levels during mitosis and early G1 phase. Unlike known substrates, KIAA1430 levels remained stable (Figure S4F).

In interphase, a fraction of GFP-tagged KIAA1430 localized to the centrosomes, in particular the centrioles, and was largely excluded from the nucleus (Figures 4G, 4H, S4G, and S4H), while the APC/C is known to be predominantly nuclear (Kraft et al., 2003; Hubner et al., 2010). During mitosis, after nuclear envelope breakdown (NEBD), APC/C accumulates on mitotic spindles, centromeres, and centrosomes (Kraft et al., 2003; Acquaviva et al., 2004), reflecting a partially common localization with KIAA1430. Consistently, we confirmed the APC/C–KIAA1430 interaction in mitotically arrested, but not in interphase cells (Figure 4I). To functionally investigate the mitotic interaction, we used time-lapse microscopy to determine the time cells require from NEBD to the onset of anaphase as a function of APC/C activity. KIAA1430 knockdown resulted in a mild delay that was sensitive to reversine, a small molecule inhibitor of the mitotic checkpoint kinase MPS1 (Figures 4J and 4K) (Santaguida et al., 2010). These findings suggest that the depletion of KIAA1430 activates the spindle assembly checkpoint, thereby postponing the activation of the APC/C. Recent reports identified the ciliary protein hemingway as the *Drosophila* ortholog of KIAA1430 (Soulavie et al., 2014) and implicated the APC/C in regulating ciliary length and polarity (Ganner et al., 2009; Wang et al., 2014). Given that centrioles are common features of cilia and centrosomes, our data suggest that in human cells, KIAA1430 recruits a sub-fraction of the APC/C to the centrosome to facilitate mitotic progression.

These examples illustrate how the combination of three quantitative dimensions offers a unique view on the interactions of individual proteins that extends beyond their identification and facilitates their functional investigation.

We have compiled this information into an easily usable resource, provided as Data S1 and available via the IntAct database. For each of the 1,330 tagged cell lines, we present a concise, one-page summary outlining the abundance of the bait protein, the co-enrichment and confidence classification of candidate interactors along with the stoichiometry plot and the predictions of the core complexes. A reading guide is presented in Figure S5.

The Relevance of Substoichiometric Interactions

Our study revealed that interactions within obligate complexes constitute only a small minority of the interactome. We reasoned that the majority of remaining interactions should be of a functionally and conceptually different nature, as indicated by our example of the KIAA1430-APC/C interaction.

To investigate the interplay of the different types of interactions, we interrogated the chaperonin TRiC (also called CCT), which is known to act on a large number of client proteins (Hartl et al., 2011). Its core machinery of eight subunits was clearly identified as an abundant obligate complex (Figure 5A) and represents a prominent hub in our interactome dataset. Virtually all interactors co-enriched with tagged TRiC core subunits were co-chaperones, regulatory proteins of the phosducin family or proteins containing known substrate motifs (Yam et al., 2008) (Table

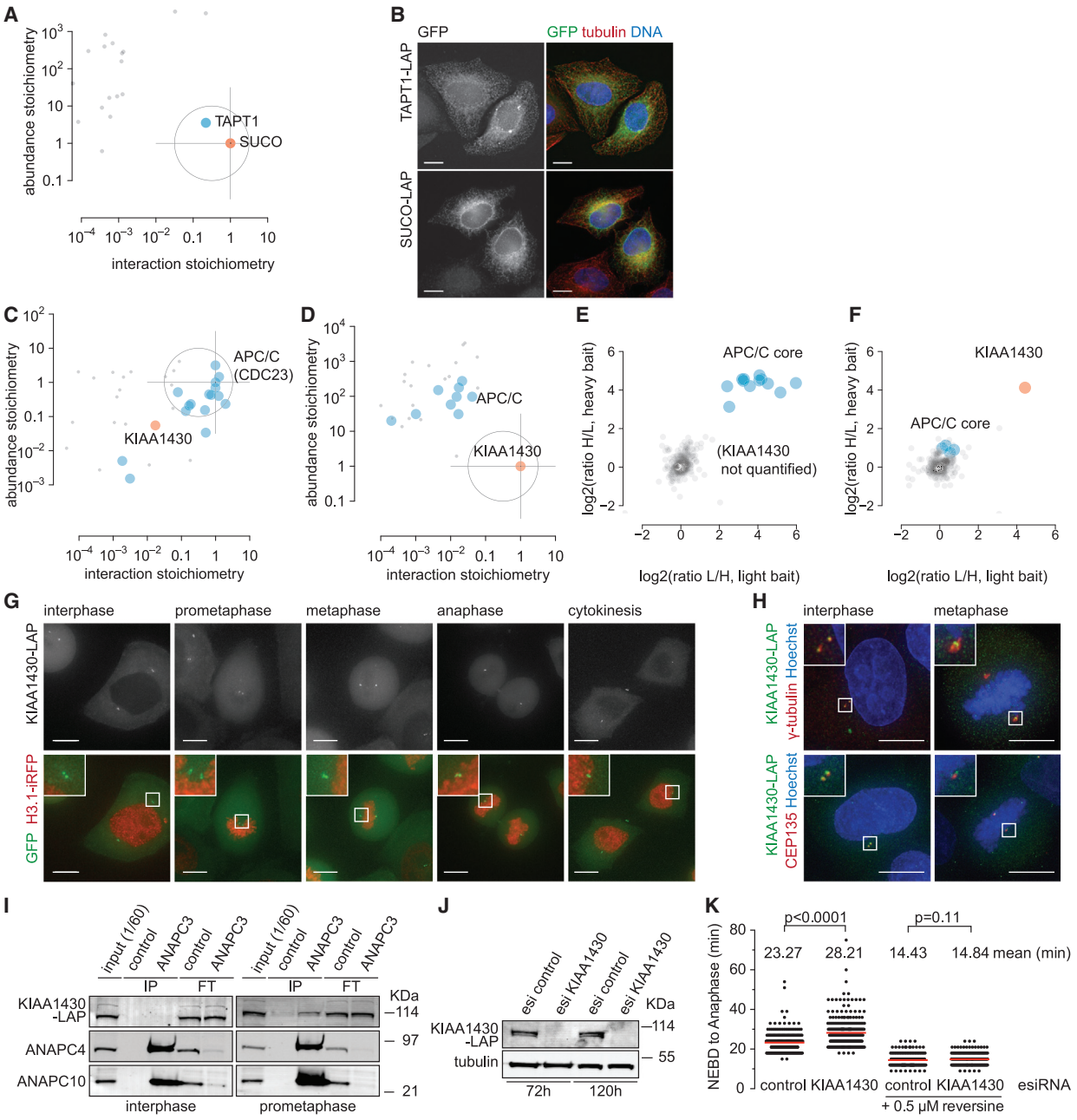


Figure 4. The TAPT1-SUCO Complex and the KIAA1430-APC/C Interaction

(A) Stoichiometry plot indicates stable TAPT1-SUCO complex.

(B) Immunofluorescence of TAPT1 and SUCO in HeLa shows ER localizations.

(C) Stoichiometry plot of APC/C interactors (bait: CDC23). Known core complex members (blue) and KIAA1430 (red) as novel substoichiometric interactor.

(D) Stoichiometry plot of KIAA1430 (red) interactors shows APC/C subunits (blue) as substoichiometric interactors.

(E) PAM-SILAC ratios plotted as medians of forward triplicate against label-swapped reverse triplicate (bait: CDC23).

(F) PAM-SILAC data using KIAA1430 as bait. Baits and stable interactors are recovered at ratios corresponding to label incorporation levels. Ratios of transient interactors are shifted toward 1:1.

(G) Maximum intensity projections of living interphase and mitotic cells expressing KIAA1430-LAP and histone 3.1-iRFP indicate that KIAA1430 localizes to centrioles.

(H) Co-localization of KIAA1430 with centrosomal marker γ-tubulin and the centriolar protein CEP135.

(I) Western blot analysis of ANAPC3 IPs and corresponding flow-throughs (FT) from interphase and mitotically arrested cells expressing KIAA1430-LAP.

(legend continued on next page)

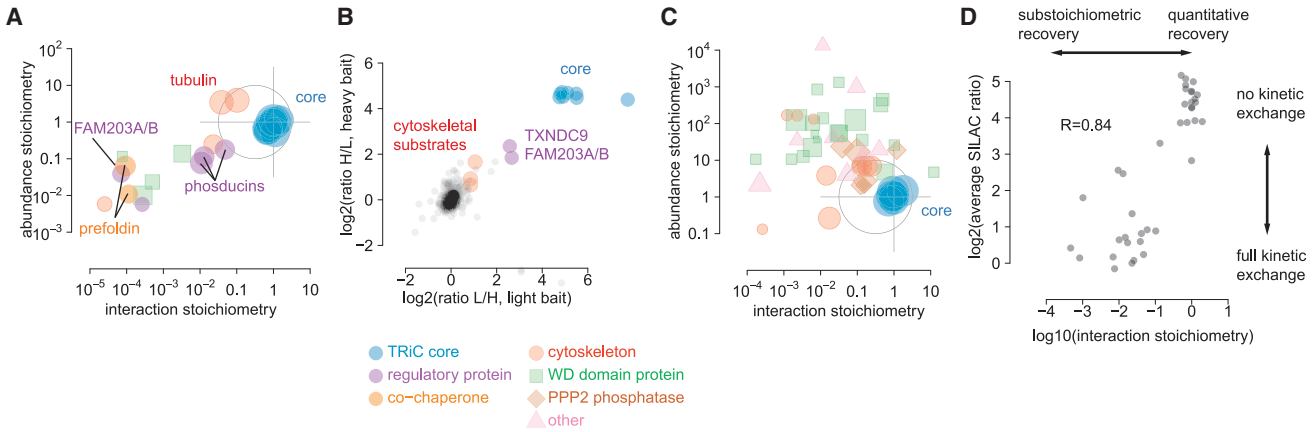


Figure 5. The TRiC Interactome Is Defined by Substoichiometric Links

(A) Stoichiometry plot of CCT3 interactors, representative of TRiC core subunits.

(B) PAM-SILAC results from the same bait protein.

(C) Reciprocal stoichiometry plot of averaged positions of the TRiC subunits from all bait pull-downs enriching at least three TRiC subunits. Symbol size indicates profile correlation. See also Tables S4 and S5.

(D) Systematic comparison of interaction stoichiometries and PAM-SILAC ratios for all interactions observed using CDC23, KIAA1430, and CCT3 as baits.

S4). Characteristic of all was a lower cellular abundance than TRiC (except for some cytoskeletal proteins) and substoichiometric recovery, classifying these interactors as distinct from the core subunits. When we performed a PAM-SILAC experiment to test for stable versus transient binding, the core complex composition that we had already established by the stoichiometry plot was confirmed, as these were all found to be stable binders (Figure 5B). Other interactors were transient, as their SILAC ratios indicated full dynamic exchange. Notable exceptions were some regulatory proteins and abundant cytoskeletal substrates, whose ratios lay between stable and fully dynamic binders. We consistently found the uncharacterized protein FAM203A/B as a substoichiometric interactor with intermediate dynamic exchange behavior. Its ortholog in *Caenorhabditis elegans* shows a cytoskeletal knockdown phenotype (Fievet et al., 2013). All of this was reminiscent of phosducin proteins, which TRiC requires to fold actin and tubulin (Hayes et al., 2011) and we therefore speculate that FAM203A/B might have a similar function.

In reciprocal interaction experiments, TRiC core complex members were co-enriched by ~5% of all bait proteins (Figure 5C; Table S5). This is in line with estimates of TRiC being involved in folding of 5%–10% of the proteome (Hartl et al., 2011). However, only some of these baits were also found in the reciprocal TRiC IPs. This asymmetry can be explained with knowledge of the underlying proteome: At 1.3 million copies of the hexadecameric complex, TRiC is much more abundant than most substrates, of which only a fraction will be in the process of folding at any given time (Table S3). Consequently, only a minute fraction of the TRiC pool will be acting on each substrate and its recovery will be "diluted" to substoichiometric levels. In the reciprocal case, however, TRiC occupies a significant fraction of the client protein population—the fraction in the state of folding—rendering the interaction more readily detectable within the dynamic range (Figure 5C).

The stoichiometry of TRiC recovery in the substrate IPs ranges from less than 10^{-3} to above 10^{-1} (Figure 5B). With TRiC substrates comprising 5%–10% of all protein molecules (a HeLa cell contains an estimated at 6×10^9 protein molecules), our stoichiometry data imply that on average 0.2%–0.4% of them are bound to the chaperone at any time. While substoichiometry may be thought to be of lower biological relevance, these interactions fulfil important functions as they connect very diverse set of protein classes. Moreover, our data also illustrate how the proteome-interactome relationship balances the amount of TRiC with the cumulative amount of its substrates.

Extrapolating from our APC/C, KIAA1430, and TRiC case studies, we investigated whether the different stoichiometric classifications of interactions carry over to other characteristics: first, we systematically compared interaction stoichiometries with dynamic exchange data for all interactions for which both orthogonal pieces of information were available (Figure 5D). There was almost perfect congruence of stoichiometric interactors with kinetically stably bound proteins and a surprisingly good overall correlation of substoichiometric recovery and the extent of dynamic exchange. This indicates that interaction stoichiometries are globally predictive of the biophysical stability of an interaction. Next, we investigated whether interaction stoichiometry is indicative of co-expression across tissues or cell types. We extracted protein abundance correlation profiles across

(J) Western analyses showing the extent of KIAA1430 depletion before and after the time-lapse analyses presented in (K).

(K) Time KIAA1430-depleted cells require to proceed from NEBD to anaphase, compared to control cells (n = 300 each); 0.5 μM reversine rescues the delay (n = 200 each). Red lines, mean. Significance according to two-tailed Mann-Whitney test. Scale bars, 10 μM.

See also Figure S4.

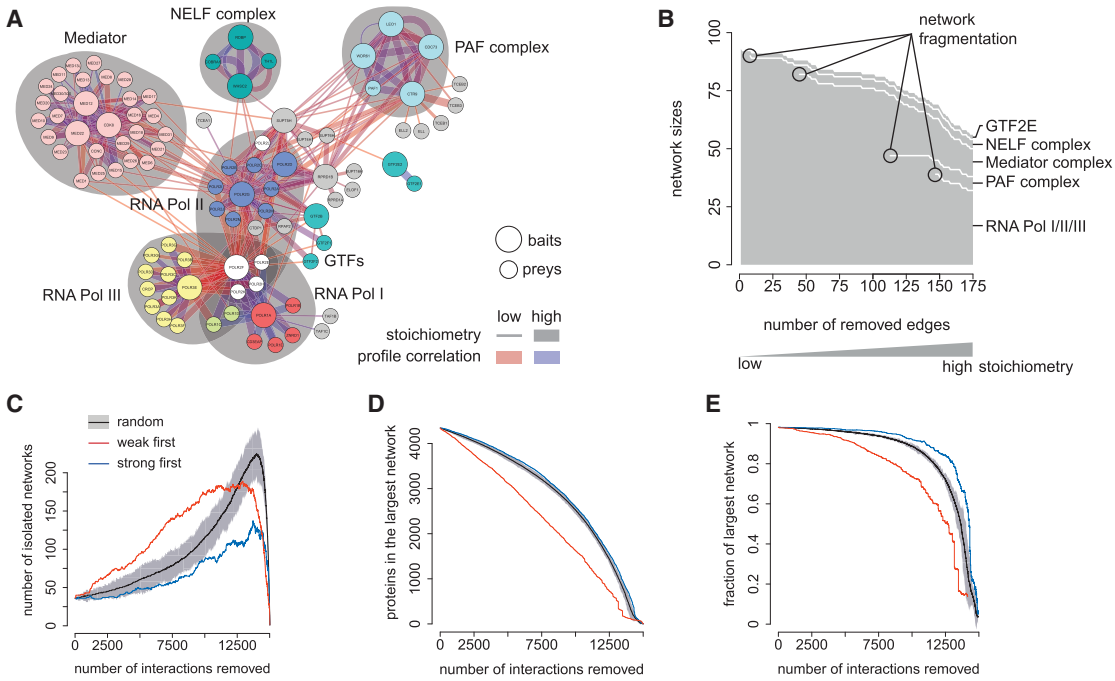


Figure 6. Strong and Weak Interactions Have Different Global Properties

(A) Sub-network of complexes surrounding RNA polymerases I/II/III. Proteins are colored by complex, edges by profile correlation, edge widths represent interaction stoichiometries.

(B) Effect of sequential removal of substoichiometric interactions on network sizes. Indicated are points where edge removal results in two fragmented sub-networks.

(C) Global network effect of random or targeted removal of interactions on the total number of isolated sub-networks.

(D) Effect on the number of proteins present in the largest entirely connected sub-network.

(E) Effect on the fraction of total connected proteins that are part of this largest sub-network.

See also Figure S6.

many tissues from a recent human proteome draft dataset (Kim et al., 2014). While co-expression coefficients scattered widely, there was still a notable relationship with interaction stoichiometry, with high-stoichiometry interactors more likely to be coherently expressed (Figure S6A). This is in agreement with earlier findings in yeast showing that members of stable complexes are enriched in co-regulated modules (Simonis et al., 2006). Conversely, substoichiometric interactions involve proteins that are not necessarily tightly co-regulated.

Finally, we tested whether interaction stoichiometry is predictive of the role of an interaction in network topology. We analyzed a sub-network of interactors surrounding RNA polymerases I, II, and III, recapitulating shared subunits and interactions with other complexes, such as general transcription factor complexes, the negative elongation factor (NELF) complex, the mediator complex, and the polymerase-associated factor (PAF) complex (Figure 6A). Sequential in silico removal of the most substoichiometric interactions from the network leads to fragmentation events, in which the individual complexes gradually lose their interconnections and emerge as individual modules (Figure 6B). Finally, the three polymerases remain internally connected via their shared subunits. Removing interactions in the reverse order does not lead to any network fragmentation, but rather results in roughly linear shrinkage of the network (Figure S6A).

Taking this approach to a global level, we probed the response of our entire network to the removal of edges according to their stoichiometry characteristics. Seminal studies on the topology of networks have shown that scale-free networks are resilient to random removal of edges, but sensitive to targeted attacks (Albert et al., 2000). Specifically, analysis of the network structure identifies the topologically most critical edges, removal of which leads to rapid network fragmentation.

In our case, we targeted edges for removal solely by their "local" interaction stoichiometry readout, agnostic to their global network roles. We removed edges sequentially, starting at either the lowest or highest interaction stoichiometry, comparing this with random removal of edges.

This revealed vastly different network responses (Figure 6C). The most substoichiometric interactions turned out to be most critical for network topology: Their preferential removal led to a rapid increase of the number of isolated network fragments, whereas removing the strongest 50% of edges hardly resulted in any network fragmentation (Figure 6C). The largest connected component, which causes the typical "hairball" appearance of large-scale networks, shrunk about linearly with removal of weak interactions (Figure 6D) and also left more proteins entirely unconnected (Figure S6A). Conversely, preferential removal of edges from the other end of the stoichiometry scale led to a

network response that increased its small-world characteristics: the largest network encompasses the vast majority of connected proteins (Figure 6E), fewer proteins are left without connections (Figure S6B) and isolated network fragments are smaller (Figure S6C). Similar patterns of network response were observed in a study analyzing mobile phone communication networks by removing the strongest versus the weakest interactions (Onnela et al., 2007). In analogy to that study, and based on our findings of the relationship between interaction stoichiometries and kinetic stabilities (Figure 5D), we propose a strong/weak terminology for interaction stoichiometries and term interactions with near-stoichiometric recovery of the prey "strong" and substoichiometric interactors "weak."

Together, our analyses show that interaction stoichiometries, which are local properties derived from single interaction experiments, predict the global behavior of the proteins involved: strong interactions are indicative of proteins that are co-regulated across cell types. In the network, they form modules of high interconnectivity, rendering the network topologically resilient to their removal. Weak interactions, on the other hand, dominate the network both in numbers and by their topologically critical role as long-range interactions between more diverse sets of proteins. As a consequence, interaction networks can be fragmented into individual, defined modules, by identifying and removing weak links. In summary, availability of interaction stoichiometries on a global scale effectively allows us to "comb" the interactome hairball, to identify modules, and visualize their interconnectedness.

DISCUSSION

Here, we have introduced a novel concept of interactome analysis. Using an efficient, low-stringent IP protocol, accurate label-free quantification of both the IPs and the complete proteome, we extracted three quantitative dimensions, all of which proved critical for characterizing protein interactions. While the first dimension identifies statistically significant interactions, the second and third dimension define their stoichiometric contexts. Earlier large-scale studies did not include or interpret these additional dimensions, in part because of the challenges involved in extracting accurate quantitative values. Moreover, past studies often employed overexpression of bait proteins, precluding meaningful stoichiometry readout (Gibson et al., 2013), and near-complete proteome coverage was also often not attainable.

Finding stable protein complexes is usually a major goal of interactomics studies. We showed that obligate protein complexes feature a unique signature of balanced stoichiometries—an infrequent occurrence among the multitude of interactions. Such a signature led us to discover the TAPT1-SUCO complex in the ER membrane. This complex ties together a body of available evidence, including knockout phenotypes of both TAPT1 and SUCO and genetic interactions of their yeast orthologs. As a representative of the majority of weaker, non-obligate interactions, we characterized the binding of KIAA1430 to the APC/C, suggesting that low interaction stoichiometries are the result of an interaction that is limited to centrioles in mitotic cells and biophysically weaker than interactions between APC/C core members. Furthermore, our stoichiometry-based classification subdivided the interactome of the TRiC chaperonin into obligate core complex subunits, regulatory interactors, and a large number of substrates. We found that lack of reciprocal verification can be indicative of an inherently asymmetric nature of biologically relevant interactions, particularly outside obligate core complexes. This example also illustrates how the observed interactome is shaped by protein abundances and, conversely, implies overall regulation of protein abundances by protein interactions. Therefore, the interactome always has to be interpreted as a function of the underlying proteome.

We have shown that interaction stoichiometries generally correlate with the biophysical stability of an interaction. Weak interactions have frequently gone undetected in interactome studies and may be thought to be less important; nevertheless they are crucial features of networks in general and social networks in particular (Granovetter, 1973; Csermely, 2006). Our study directly and quantitatively demonstrates the predominance of weak interactions in the protein interactome. MS-based methods cover more than four orders of magnitude of interaction stoichiometry (Collins et al., 2013), and our low-stringency biochemical workflow ideally harnesses this sensitivity. However, substoichiometric interactions involving low abundance preys can still be challenging to detect (Figures S3D and S3E). Therefore, the prevalence of weak interactions is likely to be even more pronounced and their relevance vastly underappreciated.

Previous studies typically counted all interactions as equal, once they had been accepted based on their statistical parameters or scores. Therefore, the roles of individual interactions had to be predicted from prior knowledge or from global network properties. Highly connected proteins were described as interaction hubs, regions of high clustering coefficients with many shared pathway annotations were characterized as complexes (Collins et al., 2007; Hart et al., 2007), and weak interactions were inferred from weaker connectivity patterns (Malovannaya et al., 2011). However, limited coverage of the interactome is a confounding factor for such strategies.

In contrast, we here have shown directly that local stoichiometry data reflect global network topological properties of interactions, setting the stage for quantitative network analysis from the ground up.

Substoichiometric interactions form the "glue" that holds the cellular network together—as shown specifically for the RNA polymerase network and globally for the entire network—and are hence critical for network structure. This property, which may seem counterintuitive at first, prompted us to propose interaction stoichiometry as a measure of interaction strength. Of note, a range of underlying mechanisms can cause a weak interaction according to this terminology, for instance low biophysical affinity, high kinetic exchange rates, limited spatiotemporal overlap of interactors, or indirect interactions that are bridged via other biomolecules, all of which may result in a substoichiometric readout. If such weak links are removed from the network, it collapses into defined modules that are tightly interconnected by the remaining strong links. Translated into biological terms, stable complexes would remain in isolation, but without weak links, they would not be able to connect to each other or to transient, dynamic regulators.

A major contribution of this study lies in the characterization of the interactomes surrounding more than 1,100 different baits, which together cover a large part of the expressed proteome with more than 28,000 interactions. We present our results in an accessible format that can be easily mined and interpreted by non-specialists. Our resource of mammalian cell lines expressing GFP-tagged proteins under endogenous control can be employed for other studies (e.g., focusing on subcellular localization or functional characterization of individual proteins). The interaction data validate these cell lines for such uses and the use of mouse orthologs as surrogates. They also imply remarkably similar protein interactomes between human and mouse.

We approach saturation with respect to the number of proteins that can be covered (Figure S6D), but observe only part of the entire interactome directly, which our data predicts to encompass between 80,000 and 180,000 detectable interactions in HeLa (Figure S6E). Our additional quantitative dimensions may prove helpful for increasing interactome coverage in silico, for example, by selective matrix expansion (Seebacher and Gavin, 2011). Given its usefulness in interpreting interaction data, the stoichiometry readout developed here can become a general basis for future interactome studies and for the analysis of interactome dynamics, which will manifest foremost as quantitative alteration of occupancies rather than qualitative gain or loss of interactors.

EXPERIMENTAL PROCEDURES

Cell Culture

HeLa Kyoto cell lines expressing N- or C-terminally tagged proteins from BAC transgenes were generated, cultured, and imaged as previously described (Poser et al., 2008). Tags are based on the "localization and affinity purification" (LAP) tag, consisting of GFP and a functionalized linker. All BAC cell lines and tag sequences are listed in Table S1 along with proteome and interactome metadata on the bait proteins. Cells were grown to near-confluency on two 15-cm cell culture dishes per interaction experiment, detached with Accutase, and snap frozen. Three replicates were harvested in at least two different passages.

Affinity Purification and Mass Spectrometry

Cell pellets were lysed and subjected to affinity purification on a robotic system, followed by single-shot mass spectrometric analysis on an Orbitrap instrument (Hubner et al., 2010). We processed triplicates separately on different days and carried out MS-analyses in randomized order over the course of weeks to months.

Whole Proteome Measurements

HeLa cells were lysed in guanidinium chloride lysis buffer and digested sequentially with LysC and trypsin as described (Kulak et al., 2014). Peptides were desalted on stacked C18 reverse phase (Waters Sep-Pak) and strong cation exchange cartridges and eluted using 70% acetonitrile. Pooled eluates were separated into six fractions on strong anion exchange (SAX) StageTips (Wiśniewski et al., 2010). MS measurements were performed in three replicates on a quadrupole Orbitrap mass spectrometer (Kulak et al., 2014).

Data Processing

Raw files were processed with MaxQuant (Cox and Mann, 2008) (version 1.3.9.10) in several sets, each containing ~600 randomly assigned AP-MS runs and the HeLa proteome fractions. Tandem mass spectrometry (MS/MS) spectra were searched against a modified version of the November 2012 release of the UniProt complete human proteome sequence database. For each bait protein expressed from a mouse BAC locus, the human sequence in the fasta file was concatenated with the mouse sequence (unless identical). Human identifiers were used for mapping purposes. We used MaxLFQ, MaxQuant's label-free quantification (LFQ) algorithm to calculate protein intensity profiles across samples (Cox et al., 2014). We required one ratio count for each pairwise comparison step and activated the FastLFQ setting with two minimum and two average comparisons to enable the normalization of large datasets in manageable computing time.

Detection of Protein Interactions

Protein identifications were filtered, removing hits to the reverse decoy database as well as proteins only identified by modified peptides. We required that each protein be quantified in all replicates from the AP-MS samples of at least one cell line. Protein LFQ intensities were logarithmized and missing values imputed by values simulating noise around the detection limit. For each protein, a non-parametric method was used to select a subset of samples that provide a distribution of background intensities for this protein (Supplemental Experimental Procedures). This subset was used first to normalize all protein intensities to represent relative enrichment and then to serve as the control group for a two-tailed Welch's t test. Specific outliers in the volcano plots of logarithmized p values against enrichments were determined by an approach making use of the asymmetry in the outlier population (Figures S1E and S1F). We used two cut-offs of different stringencies, representing 1% and 5% of enrichment false discovery rate (FDR), respectively. Correlation coefficients between the intensity profiles of interacting proteins were calculated as additional quality parameters (Keilhauer et al., 2015). Enrichment FDR (classes A–C) and profile correlation (modifier + or –) define the confidence class of an interaction (Figure S1G).

Interaction Stoichiometries and Cellular Copy Numbers

Estimating interaction stoichiometries requires the comparison of the amounts of different proteins relative to each other in one IP. We first subtracted the median intensity across all samples to account for background binding. We then divided LFQ intensities by the number of theoretically observable peptides for this protein (Schwanhäusser et al., 2011). Finally, we expressed stoichiometries relative to the bait protein. Cellular copy numbers and abundances were calculated using a similar approach (Wiśniewski et al., 2014) on the whole proteome data and brought to absolute scale by normalization to a total protein amount of 200 pg in a cell volume of 1 pl for a HeLa cell.

Network Analyses

Network analyses were performed based on the data listed in Table S2. For the purpose of counting unique interactions and for the histogram of the numbers of interactors, we regarded interactions as non-directional, flattened multiple protein groups mapping to the same gene name and to the most abundant isoform and considered interactions found multiple times only once. For network perturbation analyses, we selected all non-self-interactions of confidence classes A+, A, and B+ and assembled them into graphs. We then removed edges sequentially according to their interaction stoichiometry readout. Prey-bait combinations discovered multiple times were treated as separate edges. Once a protein had lost all its edges, it was removed. As control, we deleted edges randomly and represented the median of 100 random repetitions and represent the scatter as the first or third quartile ±1.5 interquartile ranges.

ACCESSION NUMBERS

The accession number for the protein interaction data reported in this paper, submitted through IntAct (Orchard et al., 2014), to the IMEx Consortium (http://www.imexconsortium.org): IM-24272. The accession numbers for the raw mass spectrometric data, deposited via PRIDE (Vizcaino et al., 2013) to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org): PXD002815 and PXD002829.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, five tables, and one data set and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.053.

AUTHOR CONTRIBUTIONS

M.Y.H. did the experiments, conceived and implemented the bioinformatics methods, and analyzed and interpreted the data. N.C.H. developed the QUBIC pipeline in a high throughput format. J.C. developed the MaxQuant software modules for label-free quantification of very large datasets and contributed to data analysis. N.N. provided the whole proteome data. I.P. and Y.T. generated the BAC-GFP cell lines. I.P. performed fluorescence microscopy and western blot analyses. I.A.G., I.W., and J.M. generated the data shown in Figures 4G–4K. J.M. and F.B. conceived, supervised, and interpreted the experiments in Figures 4G–4K. M.M. and A.A.H. conceived the study and supervised the experiments. M.Y.H. and M.M. wrote the manuscript.

ACKNOWLEDGMENTS

This work was performed within the project framework of the German medical genome research funded by the Federal Ministry of Education and Research (FKZ01GS0861, DiGtoP). We acknowledge funding from the Max Planck Society and the European Commission's Sixth and Seventh Framework Programs (FP6-LSHG-CT-2004-503464, MitoCheck; FP7-HEALTH-F4-2008-201648, PROSPECTS; FP7-HEALTH-2009-241548, MitoSys). J.M. is supported by the German Research Foundation (DFG) (Emmy Noether, MA 5831/1-1), I.A.G. is a member of the DIGS-BB PhD program. Antibodies for BAC validation were a gift from C. Stadler and E. Lundberg. We thank M. Leuschner, A. Ssykor, M. Augsburg, A. Schwager, M.T. Pham, G. Sowa, K. Mayr, I. Paron, B. Splettstößer, D. Vogg, B. Chatterjee, M. Grötzinger, and S. Kroiß for technical assistance. C. Schaab, S. Schloissnig, P. Porras, and M. Oroshi provided bioinformatics support. We thank M. Sarov, M. Seiler, M.C. Bassik, S. Pinkert, A. Bracher, H.C. Eberl, T. Viturawong, E.C. Keilhauer, A. Hrle, G. Pichler, N.A. Kulak, and M. Räschle for helpful discussions.

Received: February 5, 2015
Revised: July 6, 2015
Accepted: September 17, 2015
Published: October 22, 2015

REFERENCES

Acquaviva, C., Herzog, F., Kraft, C., and Pines, J. (2004). The anaphase promoting complex/cyclosome is recruited to centromeres by the spindle assembly checkpoint. Nat. Cell Biol. *6*, 892–898.

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378–382.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. *404*, 939–965.

Bassik, M.C., Kampmann, M., Lebbink, R.J., Wang, S., Hein, M.Y., Poser, I., Weibezahn, J., Horlbeck, M.A., Chen, S., Mann, M., et al. (2013). A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. Cell *152*, 909–922.

Beck, M., Claassen, M., and Aebersold, R. (2011). Comprehensive proteomics. Curr. Opin. Biotechnol. *22*, 3–8.

Collins, S.R., Kemmeren, P., Zhao, X.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S., and Krogan, N.J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol. Cell. Proteomics *6*, 439–450.

Collins, B.C., Gillet, L.C., Rosenberger, G., Röst, H.L., Vichalkovski, A., Gstaiger, M., and Aebersold, R. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nat. Methods *10*, 1246–1253.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics *13*, 2513–2526.

Csermely, P. (2006). Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks, First Edition (Springer).

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol. *3*, 89.

Fievet, B.T., Rodriguez, J., Naganathan, S., Lee, C., Zeiser, E., Ishidate, T., Shirayama, M., Grill, S., and Ahringer, J. (2013). Systematic genetic interaction screens uncover cell polarity regulators and functional redundancy. Nat. Cell Biol. *15*, 103–112.

Friederichs, J.M., Gardner, J.M., Smoyer, C.J., Whetstine, C.R., Gogol, M., Slaughter, B.D., and Jaspersen, S.L. (2012). Genetic analysis of Mps3 SUN domain mutants in Saccharomyces cerevisiae reveals an interaction with the SUN-like protein Slp1. G3 (Bethesda) *2*, 1703–1718.

Ganner, A., Lienkamp, S., Schäfer, T., Romaker, D., Wegierski, T., Park, T.J., Spreitzer, S., Simons, M., Gloy, J., Kim, E., et al. (2009). Regulation of ciliary polarity by the APC/C. Proc. Natl. Acad. Sci. USA *106*, 17799–17804.

Gibson, T.J., Seiler, M., and Veitia, R.A. (2013). The transience of transient overexpression. Nat. Methods *10*, 715–721.

Gingras, A.C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. *8*, 645–654.

Granovetter, M.S. (1973). The Strength of Weak Ties. Am. J. Sociol. *78*, 1360–1380.

Hart, G.T., Lee, I., and Marcotte, E.R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics *8*, 236.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature *475*, 324–332.

Hauri, S., Wepf, A., van Drogen, A., Varjosalo, M., Tapon, N., Aebersold, R., and Gstaiger, M. (2013). Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. Mol. Syst. Biol. *9*, 713.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., et al. (2012). A census of human soluble protein complexes. Cell *150*, 1068–1081.

Hayes, N.V.L., Jossé, L., Smales, C.M., and Carden, M.J. (2011). Modulation of phosducin-like protein 3 (PhLP3) levels promotes cytoskeletal remodelling in a MAPK and RhoA-dependent manner. PLoS ONE *6*, e28271.

Hein, M.Y., Sharma, K., Cox, J., and Mann, M. (2013). Proteomic Analysis of Cellular Systems. In Handbook of Systems Biology, *Chapter 1*, Walhout A.J.M., Vidal M., and Dekker J., eds. (Academic Press), pp. 3–25.

Howell, G.R., Shindo, M., Murray, S., Gridley, T., Wilson, L.A., and Schimenti, J.C. (2007). Mutation of a ubiquitously expressed mouse transmembrane protein (Tapt1) causes specific skeletal homeotic transformations. Genetics *175*, 699–707.

Hubner, N.C., Bird, A.W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J. Cell Biol. *189*, 739–754.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Hutchins, J.R., Toyoda, Y., Hegemann, B., Poser, I., Hériché, J.K., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., et al. (2010). Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science *328*, 593–599.

Jonikas, M.C., Collins, S.R., Denic, V., Oh, E., Quan, E.M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J.S., and Schuldiner, M. (2009). Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science *323*, 1693–1697.

Kachroo, A.H., Laurent, J.M., Yellman, C.M., Meyer, A.G., Wilke, C.O., and Marcotte, E.M. (2015). Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science *348*, 921–925.

Keilhauer, E.C., Hein, M.Y., and Mann, M. (2015). Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). Mol. Cell. Proteomics *14*, 120–135.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kittler, R., Pelletier, L., Ma, C., Poser, I., Fischer, S., Hyman, A.A., and Buchholz, F. (2005). RNA interference rescue by bacterial artificial chromosome transgenesis in mammalian tissue culture cells. Proc. Natl. Acad. Sci. USA *102*, 2396–2401.

Kraft, C., Herzog, F., Gieffers, C., Mechtler, K., Hagting, A., Pines, J., and Peters, J.M. (2003). Mitotic regulation of the human anaphase-promoting complex by phosphorylation. EMBO J. *22*, 6598–6609.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods *9*, 907–909.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Maliga, Z., Junqueira, M., Toyoda, Y., Ettinger, A., Mora-Bermúdez, F., Klemm, R.W., Vasilj, A., Guhr, E., Ibarlucea-Benitez, I., Poser, I., et al. (2013). A genomic toolkit to investigate kinesin and myosin motor function in cells. Nat. Cell Biol. *15*, 325–334.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the human endogenous coregulator complexome. Cell *145*, 787–799.

Mann, M., Kulak, N.A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. Mol. Cell *49*, 583–590.

Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.P., St-Denis, N.A., Li, T., Miteva, Y.V., Hauri, S., Sardiu, M.E., Low, T.Y., et al. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods *10*, 730–736.

Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.L. (2007). Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. USA *104*, 7332–7336.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363.

Phillips-Krawczak, C.A., Singla, A., Starokadomskyy, P., Deng, Z., Osborne, D.G., Li, H., Dick, C.J., Gomez, T.S., Koenecke, M., Zhang, J.S., et al. (2015). COMMD1 is linked to the WASH complex and regulates endosomal trafficking of the copper transporter ATP7A. Mol. Biol. Cell *26*, 91–103.

Poser, I., Sarov, M., Hutchins, J.R., Hériché, J.K., Toyoda, Y., Pozniakovsky, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A.W., et al. (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat. Methods *5*, 409–415.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173–1178.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Santaguida, S., Tighe, A., D'Alise, A.M., Taylor, S.S., and Musacchio, A. (2010). Dissecting the role of MPS1 in chromosome biorientation and the spindle checkpoint through the small molecule inhibitor reversine. J. Cell Biol. *190*, 73–87.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Seebacher, J., and Gavin, A.C. (2011). SnapShot: Protein-protein interaction networks. Cell *144*, 1000.

Simonis, N., Gonze, D., Orsi, C., van Helden, J., and Wodak, S.J. (2006). Modularity of the transcriptional response of protein complexes in yeast. J. Mol. Biol. *363*, 589–610.

Smits, A.H., Jansen, P.W., Poser, I., Hyman, A.A., and Vermeulen, M. (2012). Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. Nucleic Acids Res. *41*, e28.

Sohaskey, M.L., Jiang, Y., Zhao, J.J., Mohr, A., Roemer, F., and Harland, R.M. (2010). Osteopotentia regulates osteoblast maturation, bone formation, and skeletal integrity in mice. J. Cell Biol. *189*, 511–525.

Soulavie, F., Piepenbrock, D., Thomas, J., Vieillard, J., Duteyrat, J.L., Cortier, E., Laurençon, A., Göpfert, M.C., and Durand, B. (2014). hemingway is required for sperm flagella assembly and ciliary motility in Drosophila. Mol. Biol. Cell *25*, 1276–1286.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957–968.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

Vizcaino, J.A., Cote, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. *41*, D1063–D1069.

Wang, X., and Huang, L. (2008). Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. Mol. Cell. Proteomics *7*, 46–57.

Wang, W., Wu, T., and Kirschner, M.W. (2014). The master cell cycle regulator APC-Cdc20 regulates ciliary length and disassembly of the primary cilium. eLife *3*, e03083.

Wepf, A., Glatter, T., Schmidt, A., Aebersold, R., and Gstaiger, M. (2009). Quantitative interaction proteomics using mass spectrometry. Nat. Methods *6*, 203–205.

Wiśniewski, J.R., Nagaraj, N., Zougman, A., Gnad, F., and Mann, M. (2010). Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. J. Proteome Res. *9*, 3280–3289.

Wiśniewski, J.R., Hein, M.Y., Cox, J., and Mann, M. (2014). A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. Mol. Cell. Proteomics *13*, 3497–3506.

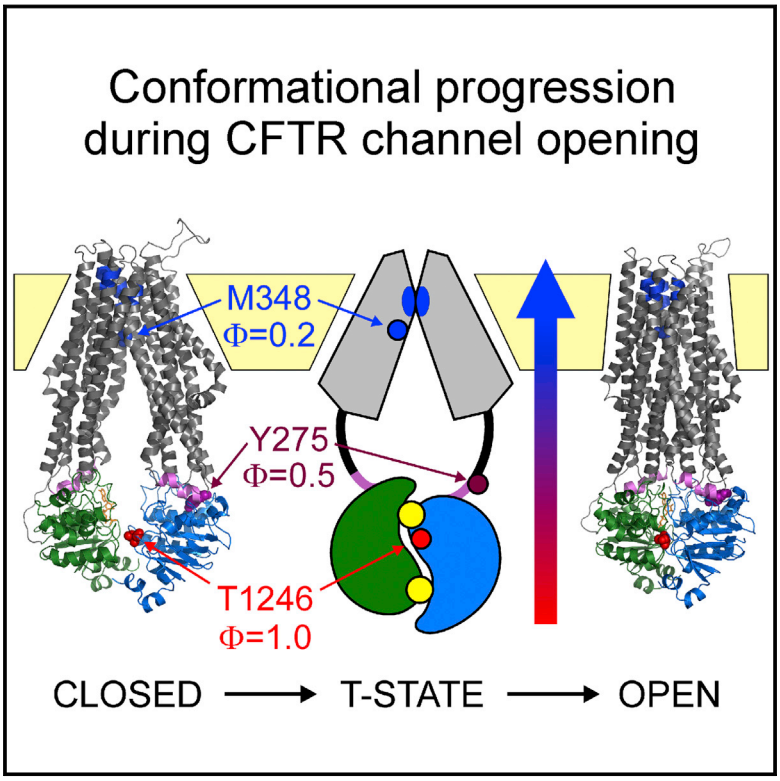
Yam, A.Y., Xia, Y., Lin, H.T., Burlingame, A., Gerstein, M., and Frydman, J. (2008). Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. Nat. Struct. Mol. Biol. *15*, 1255–1262.

Article

Cell

Timing of CFTR Pore Opening and Structure of Its Transition State

Graphical Abstract



Authors

Ben Sorum, Dávid Czégé, László Csanády

Correspondence

csanady.laszlo@med.semmelweis-univ.hu

In Brief

Characterization of the transition state during opening of the CFTR channel pore reveals a "wave" of conformational changes through the complex and identifies strain at the interface where the prevalent cystic fibrosis mutation occurs, thereby informing its role in disease-related defects in channel function.

Highlights

- In CFTR Cl^- channels, opening motion spreads from cytoplasmic to extracellular regions
- In the transition state, the nucleotide binding domains are already tightly dimerized
- In the transition state, the ion conducting pore is still closed
- Augmented strain at transmission interface causes functional defect of disease mutant

Sorum et al., 2015, Cell *163*, 724–733
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.052

CellPress

Cell

Article

Timing of CFTR Pore Opening and Structure of Its Transition State

Ben Sorum,[1,2] Dávid Czégé,[2] and László Csanády[1,2,*]
[1]Department of Medical Biochemistry
[2]MTA-SE Ion Channel Research Group
Semmelweis University, Tűzoltó u. 37-47, Budapest 1094, Hungary
*Correspondence: csanady.laszlo@med.semmelweis-univ.hu
http://dx.doi.org/10.1016/j.cell.2015.09.052

SUMMARY

In CFTR, the chloride ion channel mutated in cystic fibrosis (CF) patients, pore opening is coupled to ATP-binding-induced dimerization of two cytosolic nucleotide binding domains (NBDs) and closure to dimer disruption following ATP hydrolysis. CFTR opening rate, unusually slow because of its high-energy transition state, is further slowed by CF mutation ΔF508. Here, we exploit equilibrium gating of hydrolysis-deficient CFTR mutant D1370N and apply rate-equilibrium free-energy relationship analysis to estimate relative timing of opening movements in distinct protein regions. We find clear directionality of motion along the longitudinal protein axis and identify an opening transition-state structure with the NBD dimer formed but the pore still closed. Thus, strain at the NBD/pore-domain interface, the ΔF508 mutation locus, underlies the energetic barrier for opening. Our findings suggest a therapeutic opportunity to stabilize this transition-state structure pharmacologically in ΔF508-CFTR to correct its opening defect, an essential step toward restoring CFTR function.

INTRODUCTION

The cystic fibrosis (CF) transmembrane conductance regulator (CFTR) is the chloride ion channel mutated in patients suffering from CF, a devastating multiorgan disease (O'Sullivan and Freedman, 2009). The majority (~90%) of CF patients carry at least one allele with a deletion of phenylalanine 508 (ΔF508). The ΔF508 mutation severely impairs both surface expression (Cheng et al., 1990) and chloride channel function (Miki et al., 2010) of CFTR. Even if efforts to promote trafficking to the surface prove successful, understanding the molecular mechanism of the functional defect in ΔF508 CFTR will still be essential for its correction in CF patients.

CFTR belongs to the family of ATP binding cassette (ABC) transporters (Riordan et al., 1989), which are built from two homologous halves, each comprising a transmembrane domain (TMD) (Figures 1A and 1B, gray) followed by a cytosolic nucleotide binding domain (NBD) (Figures 1A and 1B, green and blue). In CFTR, these two halves (TMD1-NBD1 and TMD2-NBD2) are linked by the unique cytosolic regulatory (R) domain (Figure 1A, top, magenta), a target for phosphorylation by cAMP-dependent protein kinase (PKA) (Riordan et al., 1989); R-domain phosphorylation is a prerequisite for CFTR chloride channel activity (Tabcharani et al., 1991).

Opening and closing (gating) of the CFTR chloride ion pore, formed by its TMDs, is coupled to a conserved ATP binding/hydrolysis cycle at the NBDs (Figure 1B). In ABC proteins, ATP binding triggers association of the two NBDs into a stable head-to-tail dimer that occludes two molecules of ATP (Figure 1B, yellow circles) at the interface (Smith et al., 2002). By forming strong interactions with conserved residues of both the Walker motifs in the head of one NBD and the signature sequence in the tail of the opposing NBD, these ATP molecules act as molecular glue that ties the NBDs together: prompt dimer disruption therefore requires ATP hydrolysis (Moody et al., 2002). In CFTR, only the composite binding site formed by Walker A and B motifs of NBD2 and the signature sequence of NBD1 (site 2; Figure 1B, upper site) is catalytically active; the other interfacial binding site (site 1; Figure 1B, lower site) is degenerate and keeps ATP bound and unhydrolyzed throughout several NBD dimerization cycles (Aleksandrov et al., 2002; Basso et al., 2003). In ABC exporters, NBD dimer formation flips the TMDs to an outward-facing orientation, while dimer disruption following ATP hydrolysis and substrate release resets them to inward-facing; NBD-to-TMD signal transmission is mediated by an interface that includes four short "coupling helices" (CH1–4) (Locher, 2009) in TMD intracellular loops (Figure 1A, violet loops). In CFTR, NBD dimer formation initiates a burst of pore openings interrupted by brief closures, while dimer dissociation terminates the burst and returns the TMDs into a long-lasting nonconducting (interburst) state (Vergani et al., 2005). Functional studies confirm that in the bursting ("open") state CFTR's TMDs resemble the outward-facing, whereas in the interburst ("closed") state they resemble the inward-facing conformation of ABC exporter TMDs (Bai et al., 2011; Cui et al., 2014; Wang et al., 2014a). For wild-type (WT) CFTR gating is a unidirectional cycle: most openings are terminated by ATP hydrolysis (Figure 1B; step $O_1 \rightarrow O_2$) rather than by far slower non-hydrolytic closure (Figure 1B; step $O_1 \rightarrow C_1$; rate k_{OC}) (Csanády et al., 2010).

The major functional defect of ΔF508 CFTR is a >40-fold reduction in opening rate (Miki et al., 2010; Kopeikin et al., 2014), which reflects destabilization—relative to the closed

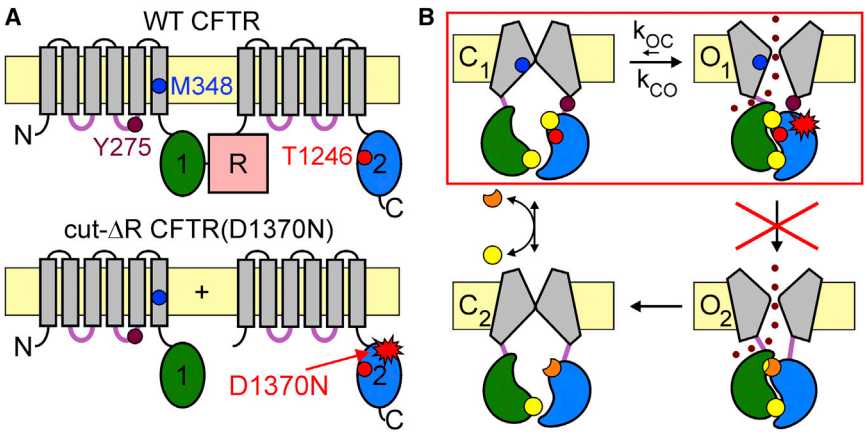


Figure 1. Choice of a Suitable Background Construct for REFER Analysis

(A) Domain organizations of WT (top) and cut-ΔR(D1370N) (bottom) CFTR: TMDs (gray), intracellular loops containing coupling helices (light violet), NBD1 (green), NBD2 (blue), R domain (magenta), membrane (yellow). Colored circles identify target positions. The D1370N mutation in NBD2 is depicted by a red star.

(B) Cartoon representation of the CFTR gating cycle. ATP-bound closed channels (C_1) open to a prehydrolytic open state (O_1). During most openings ATP is hydrolyzed at composite site 2 (state O_2), prompting NBD dimer dissociation and pore closure (to state C_2) followed by ADP-ATP exchange (to state C_1). Color coding as in (A). ATP, yellow circles; ADP, orange crescents. The D1370N mutation abrogates ATP hydrolysis (red cross) and confines gating in saturating ATP to a simple $C_1 \leftrightarrow O_1$ equilibrium (red box). Target positions color coded as in (A).

See also Figures S1 and S2.

state—of the transition state for channel opening (step $C_1 \rightarrow O_1$; Figure 1B; rate k_{CO}). Insight into the dynamics of this opening conformational change and the structure of its transition state would be a key step forward in understanding the molecular mechanism of the ΔF508 gating defect.

Transition states, which determine the rates of functionally relevant conformational movements, are the highest-energy, shortest-lived conformations of proteins. For instance, for ion channels closed ↔ open conformational transitions are so fast that they appear as single steps even in the highest time-resolution recordings, implying that the time the channel protein spends in the transition state itself is on the sub-microsecond scale—in contrast to the long (milliseconds-seconds) intervals spent in comparatively stable open and closed ground states observable in single-channel recordings (Figures 2A, 3A, 4A, and 5A). Intractable by standard structural biological approaches, transition-state structures can be studied using rate-equilibrium free-energy relationship (REFER) analysis, which reports on the relative timing of movements in selected protein regions during a conformational transition, such as a channel opening step (Zhou et al., 2005; Auerbach, 2007). Structural perturbations (typically point mutations) in a given channel region often change channel open probability (P_o) by affecting the open-closed free-energy difference, but the extent to which the free energy of the barrier that must be traversed in opening or closing the channel is affected depends on how early or late that region moves. A region that moves early during opening will have already approached its open-state conformation in the overall transition state: a perturbation here thus affects the stability of the transition state to an extent similar to that of the open state and so impacts only opening, but not closing, rate. In contrast, a region that moves late during opening is still near its closed-state conformation in the transition state: a perturbation here that affects open-state stability thus does not affect transition-state stability and so impacts only closing, but not opening, rate. This relative timing of motion of a given region of the channel protein is reported by its Φ value, the slope of a log-log plot of opening rate (k_{CO}) versus equilibrium constant ($K_{eq} = P_o/(1 - P_o)$) for a series of structural perturbations (Brønsted plot): a large Φ value (close to 1) indicates early movement, and a small Φ value (close to 0) indicates late movement.

REFER analysis has been extremely fruitful in mapping gating dynamics of the nicotinic acetylcholine receptor channel (Mitra et al., 2005; Purohit et al., 2007, 2013, 2015) but is applicable only to equilibrium mechanisms (Csanády, 2009), unlike that of CFTR. This drawback has so far hampered insight into the CFTR opening transition state. Here, we exploit a catalytic site mutation that abolishes ATP hydrolysis and so truncates the CFTR gating cycle to an equilibrium process. In this non-hydrolytic background, we employ the REFER technique to address the relative timing of movements within the sub-microsecond process of pore opening in three spatially distant positions distributed along the longitudinal axis (cytoplasmic to extracellular) of the CFTR protein. Our results identify a conformation-change wave with clear directionality and provide direct measurements that outline the global structure of the CFTR opening transition state.

RESULTS

Choice of a Background Construct Suitable for REFER Studies

ATP hydrolysis in ABC proteins is destroyed by mutations of the Walker B aspartate (Urbatsch et al., 1998; Hrycyna et al., 1999; Rai et al., 2006) that coordinates Mg^{2+} at each active site (Hung et al., 1998; Hopfner et al., 2000). To make CFTR gate at equilibrium, we introduced the NBD2 Walker-B mutation D1370N (Figure 1A, bottom, red star) because, among several hydrolysis-disrupting mutations tested, D1370N only slightly reduces the apparent affinity for ATP, and does not prolong open bursts to an extent incompatible with single-channel gating analysis (Csanády et al., 2010).

PKA- and ATP-dependent regulation of CFTR gating are intertwined, and the mechanism of R-domain action is poorly understood: evidence exists for its direct interaction with both NBDs and TMDs (Wang et al., 2002; Bozoky et al., 2013). Thus,

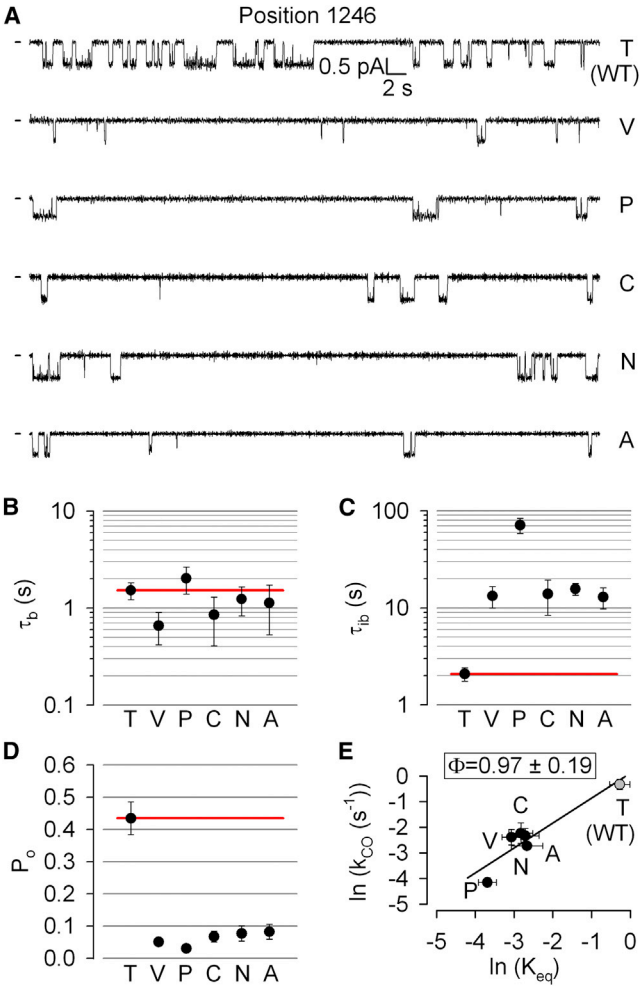


Figure 2. Timing of Motion at Position 1246 of the NBD1-NBD2 Interface

(A) Inward single-channel currents of the cut-ΔR(D1370N) CFTR background construct (top trace) and of channels bearing mutations T1246V, T1246P, T1246C, T1246N, and T1246A, respectively, in the same background. Currents were recorded at −80 mV, in symmetrical 140 mM Cl^-; dashes on the left mark zero-current level.

(B–D) Mean burst (B, τ_b) and interburst (C, τ_{ib}) durations and open probabilities (D, P_o) of the six constructs in (A). Red horizontal lines highlight the respective control values of the background construct. All data are shown as mean ± SEM (n = 4–28).

(E) Brønsted plot for position 1246. Gray symbol identifies the background construct (also in Figures 3E, 4E, and 5E). Solid line is a linear regression fit with slope Φ indicated.

See also Figures S1 and S2.

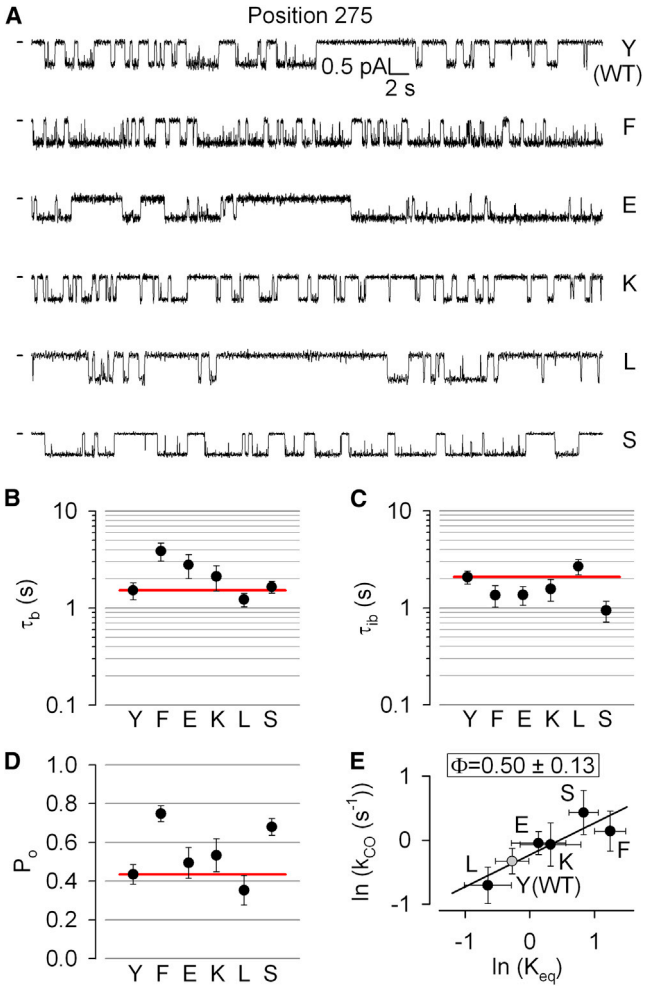


Figure 3. Timing of Motion at Position 275 of the NBD2-TMD Interface

(A) Inward single-channel currents of the cut-ΔR(D1370N) CFTR background construct (top trace) and of channels bearing mutations Y275F, Y275E, Y275K, Y275L, and Y275S, respectively, in the same background. Currents were recorded at −80 mV, in symmetrical 140 mM Cl^-; dashes on the left mark zero-current level.

(B–D) Mean burst (B, τ_b) and interburst (C, τ_{ib}) durations and open probabilities (D, P_o) of the six constructs in (A). Red horizontal lines highlight the respective control values of the background construct. All data are shown as mean ± SEM (n = 9–28).

(E) Brønsted plot for position 275. Solid line is a linear regression fit with slope Φ indicated.

See also Figures S1 and S2.

changes in gating kinetics caused by perturbations of a target position might potentially reflect altered R-domain/target position interactions, rather than energetic effects on ATP-dependent conformational transitions. Such confounding effects are absent in channels lacking the R domain: cut-ΔR channels, obtained by coexpression of TMD1-NBD1 (residues 1–633) and TMD2-NBD2 (residues 837–1480), do not require phosphorylation to be active, while ATP-dependent gating remains similar to WT (Csanády et al., 2000).

Thus, we chose cut-ΔR(D1370N) as the background construct for our REFER study (Figure 1A, bottom). Gating of cut-ΔR(D1370N) indeed proved PKA-independent but remained strictly ATP-dependent with an apparent affinity for ATP of 288 ± 27 μM (Figures S1A and S1B). Just as for WT (Winter et al., 1994; Zeltwanger et al., 1999; Csanády et al., 2000; Vergani et al., 2003), cut-ΔR (Csanády et al., 2000; Bompadre et al., 2005), and D1370N (Vergani et al., 2003) CFTR channels, mean open burst duration (τ_b) of cut-ΔR(D1370N) proved largely

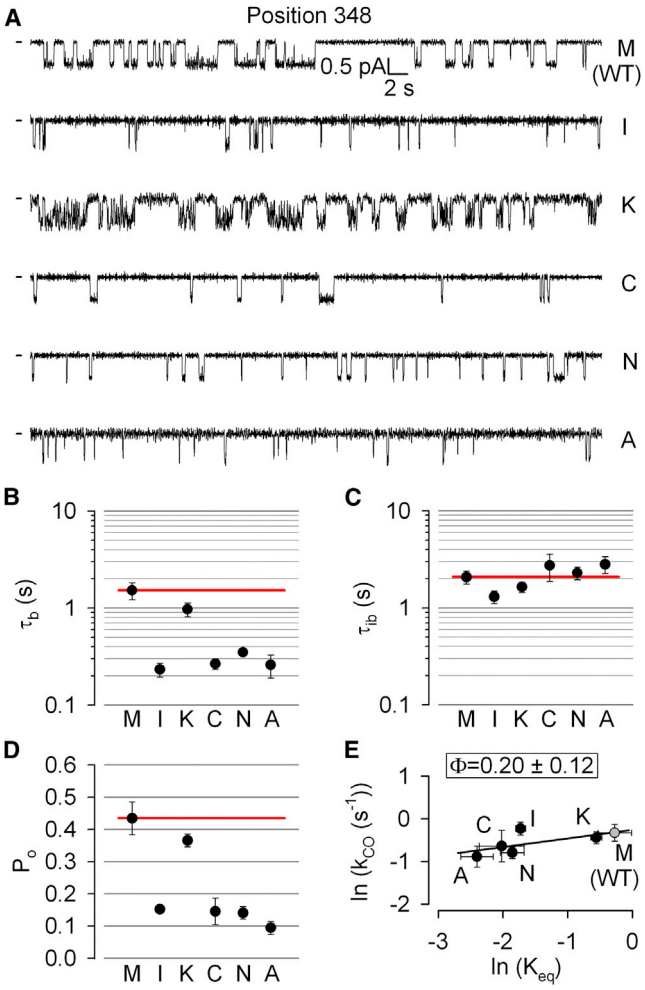


Figure 4. Timing of Motion at Position 348 in the Pore Region

(A) Inward single-channel currents of the cut-ΔR(D1370N) CFTR background construct (top trace) and of channels bearing mutations M348I, M348K, M348C, M348N, and M348A, respectively, in the same background. Currents were recorded at −80 mV, in symmetrical 140 mM Cl^-; dashes on the left mark zero-current level.

(B–D) Mean burst (B, τ_b) and interburst (C, τ_{ib}) durations and open probabilities (D, P_o) of the six constructs in (A). Red horizontal lines highlight the respective control values of the background construct. All data are shown as mean ± SEM (n = 4–28).

(E) Brønsted plot for position 348. Solid line is a linear regression fit with slope Φ indicated.

See also Figures S1 and S2.

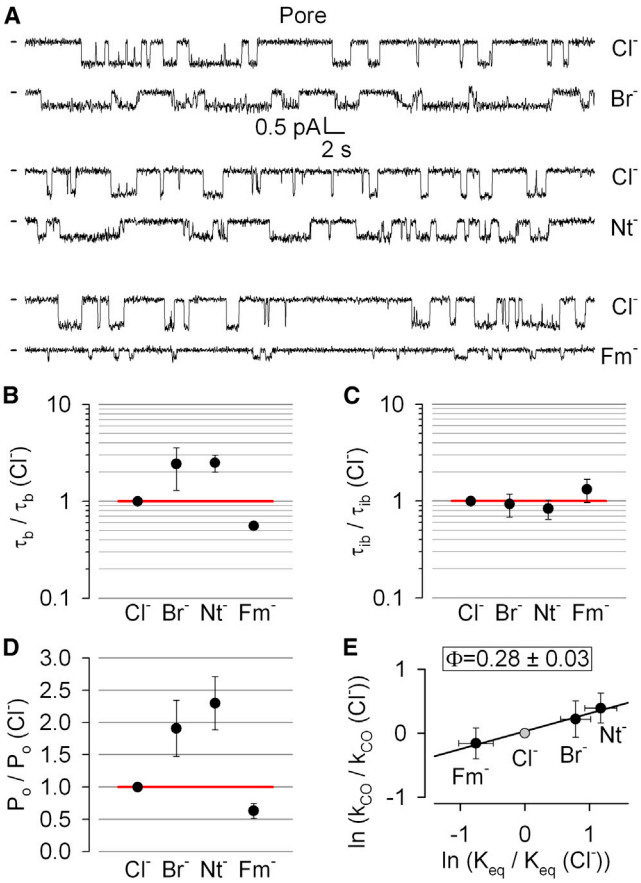


Figure 5. Timing of Motion in the Narrow Region of the Pore Studied by Anion Replacement

(A) Pairs of segments of inward single-channel current from three patches containing single cut-ΔR(D1370N) CFTR channels. Each patch was alternately exposed to bath solutions containing 140 mM of either chloride (upper segments) or a test anion (lower segments), as indicated to the right: chloride (Cl^-), bromide (Br^-), nitrate (Nt^-), formate (Fm^-). Currents in Cl^-, Br^-, and Nt^- were recorded at −80 mV, those in Fm^- at −100 mV; dashes on the left mark zero-current level.

(B–D) Mean burst (B, τ_b) and interburst (C, τ_{ib}) durations and open probabilities (D, P_o) in the presence of various test anions, each normalized to the value observed in chloride in the same patch. Red horizontal lines highlight the respective control values in chloride. All data are shown as mean ± SEM (n = 5–9).

(E) Brønsted plot for the narrow region of the pore, constructed from normalized opening rates and equilibrium constants in the presence of the four permeating anions tested. Solid line is a linear regression fit with slope Φ indicated.

See also Figure S2.

ATP-independent: the time constant of macroscopic current relaxation following ATP removal (1,342 ± 72 ms; Figures S2A and S2H), a measure of τ_b in zero ATP, was similar to steady-state τ_b of single channels in saturating (10 mM) ATP (1,526 ± 301 ms; Figures 2A and 2B). Thus, ATP-dependence of P_o reflects ATP-dependence of its mean interburst duration (τ_{ib}). Importantly, in saturating (10 mM) ATP τ_{ib} is minimal, and bursting of this background construct is reduced to a simple equilibrium ($C_1 \leftrightarrow O_1$; Figure 1B, red box; see histograms of burst and interburst durations in Figure S3) suitable for study by the REFER approach.

Timing of Movements in Composite Site 2 of the NBD1-NBD2 Interface

NBD2 Walker-A threonine 1246 makes important contributions to forming composite site 2 of the CFTR NBD1-NBD2 dimer, by contacting the γ-phosphate of ATP (PDB: 3GD7). Moreover, this interfacial residue undergoes relative movement upon NBD dimerization, as reported by interaction of its side chain across the dimer interface with that of opposing NBD1 residue R555 in open, but not closed, channels (Vergani et al., 2005). To test

timing of relative movement at this NBD interface position, we created a series of mutants by replacing the native threonine with valine, proline, cysteine, asparagine, and alanine, respectively, and characterized their gating kinetics in inside-out single-channel patches superfused with 10 mM ATP (Figure 2A). All of these perturbations dramatically reduced P_o (Figures 2A and 2D) by prolonging mean interburst duration (τ_{ib}; Figures 2A and 2C), i.e., by slowing channel opening rate ($k_{CO} = 1/\tau_{ib}$). In comparison, mean open burst durations (τ_b), and hence closing rates ($1/\tau_b$), were less affected (Figures 2A and 2B). Correspondingly, the Brønsted plot for position 1246 yielded a steep slope of $\Phi = 0.97 \pm 0.19$ (Figure 2E), indicating that this position moves very early during the pore opening conformational transition. Importantly, although mutations at position 1246 slightly reduce the affinity for ATP binding (Vergani et al., 2005), 10 mM ATP remained saturating for each of the mutants (Figure S1C, red bars), confirming that their reduced opening rates indeed reflect slowing of step $C_1 \rightarrow O_1$ (k_{CO}, Figure 1B).

Timing of Movements at the NBD2-TMD Interface

The four coupling helices at the NBD-TMD transmission interface undergo large movements between inward- and outward-facing conformations (Dawson and Locher, 2006; Hohl et al., 2012). Due to the domain swapping observed in ABC exporters, CH2 of TMD1 (residues 270–274) is in contact with NBD2 (He et al., 2008), and tyrosine 275 at the C-terminal end of CH2 is part of a conserved aromatic cluster important for NBD2-TMD interactions (Mornon et al., 2008). To test timing of motions at the NBD2-TMD transmission interface, we substituted phenylalanine, glutamate, lysine, leucine, and serine, respectively, for tyrosine 275 and studied gating of single mutant channels in 10 mM ATP (Figure 3A). Perturbations at position 275 caused modest changes in P_o but in both directions (Figures 3A and 3D). Kinetic analysis revealed a clear tendency for opposing effects on channel closing and opening rates, both contributing about equally to changes in P_o: lengthened τ_b was mostly associated with shortened τ_{ib} and (in Y275L) shortened τ_b with lengthened τ_{ib} (Figures 3B and 3C). Changes in opening rate ($1/\tau_{ib}$) again reflected changes in rate k_{CO} (Figure 1B), since 10 mM ATP remained saturating for each mutant (Figure S1C, violet bars). These coupled changes in opening and closing rates resulted in a Brønsted plot with an intermediate slope of $\Phi = 0.50 \pm 0.13$ (Figure 3E), indicating that position 275 has not yet reached its final open-like position in the opening transition state.

Timing of Movements in the Pore Region

Mutations of several pore residues were reported to affect gating (Zhang et al., 2000; Beck et al., 2008; Bai et al., 2010; Gao et al., 2013), indicating gating-related movements at these pore positions. To study the timing of such movements, we chose position 348 in transmembrane helix 6, because mutations here profoundly affected P_o without major effects on conductance (Bai et al., 2010), making it an attractive target for single-channel gating studies. To perturb position 348, we systematically replaced the native methionine with isoleucine, lysine, cysteine, asparagine, or alanine and recorded single-channel currents of the mutants in 10 mM ATP (Figure 4A), a saturating concentration for all constructs (Figure S1C, blue bars). Except for the lysine substitution, perturbations at position 348 all dramatically reduced P_o (Figures 4A and 4D), and this effect was in every case due to speeding of closing rate (reduction in τ_b, Figure 1B), with little change in opening rate ($1/\tau_{ib}$, cf., Figure 4C). Interestingly, the M348K mutation only marginally affected gating (Figures 4A–4D) but increased the affinity for pore block by ATP, as reported by pronounced flickery block of single-channel currents in 10 mM ATP (Figure 4A), a bell-shaped ATP dose-dependence of macroscopic currents (Figure S1C, second blue bar), and a current overshoot upon ATP removal from macroscopic patches reflecting rapid unblock (Figure S2F). Of note, even for M348K, the macroscopic current relaxation time constant following ATP removal (i.e., τ_b in zero ATP; Figure S2F) remained comparable to steady-state τ_b: thus, even pronounced flickery block of M348K by 10 mM ATP does not delay pore closure, consistent with earlier demonstration that the gate, located on the extracellular side, can readily close while large organic anion blockers remain bound in the intracellular vestibule (Csanády and Töröcsik, 2014). We also replaced the methionine with glutamate, but this M348E mutant could not be studied at a single-channel level due to the presence of subconductance states; however, the rate of macroscopic current relaxation upon ATP removal attested to an acceleration of M348E closing rate comparable to that of the I, C, N, and A mutants (Figures S2G and S2H). This speeding of non-hydrolytic closure (step $O_1 \rightarrow C_1$ in Figure 1B, rate k_{OC}) by perturbations at position 348, with little effect on opening rate, led to a Brønsted plot with a small slope of $\Phi = 0.20 \pm 0.12$ (Figure 4E), indicating that this pore region still resembles its closed-state conformation in the opening transition state.

Timing of Movements in the Narrow Region of the Pore Studied by Anion Substitution

Previous accessibility studies outlined a short narrow region of the pore, confined to approximately one helical turn of pore-forming transmembrane helices 1 (residues 102–106), 6 (residues 337–341), 11 (residues 1115–1118), and 12 (residues 1130–1136) (Beck et al., 2008; Fatehi and Linsdell, 2009; El Hiani and Linsdell, 2010; Bai et al., 2010; Qian et al., 2011; Wang et al., 2011; Gao et al., 2013; Wang et al., 2014b). This narrow region was shown to act as a lyotropic "selectivity filter" that provides sites of interaction (T338, S341, S1118, T1134) for permeating anions (McDonough et al., 1994; Linsdell et al., 2000; Linsdell, 2001; Zhang et al., 2000; McCarty and Zhang, 2001). Intriguingly, replacement of chloride with nitrate affects CFTR gating (Yeh et al., 2015), suggesting that interactions of permeating anions with residues lining the "filter" region of the open pore energetically contribute to open-state stability. Thus, replacement of chloride with other permeant anions might be viewed as a structural perturbation of the "selectivity filter." We therefore studied changes in the pattern of single-channel gating of our background construct cut-ΔR(D1370N) in response to sudden replacement of cytosolic chloride with nitrate, bromide, or formate. Of note, in these experiments gating in chloride and in the replacement anion could be compared within the same patch (Figure 5A): this arrangement eliminates any uncertainties about perturbation-induced fractional changes in opening rate, precise estimation of which is

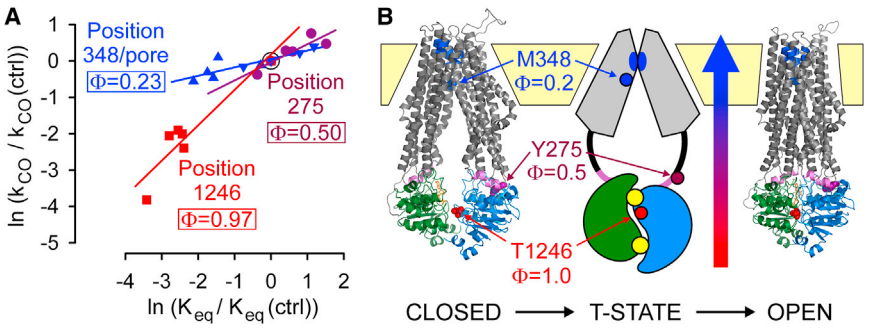


Figure 6. Opening Conformational Wave and Transition-State Structure Reported by Φ-Value Analysis

(A) Merged normalized Brønsted plot for the pore region (blue). Fitting the ensemble of the normalized data for position 348 (standing triangles) and for the anion substitution experiments (inverted triangles) by linear regression (solid blue line) yielded the indicated slope value (Φ). Brønsted plots for positions 1246 (red) and 275 (violet) are normalized versions of the plots in Figures 2E and 3E, respectively. The point representing the cut-ΔR(D1370N) background construct in chloride is highlighted by a black circle.

(B) Ribbon representation of CFTR homology models (Corradi et al., 2015) in the closed (left) and open (right) states based on (left) the inward-facing structure of TM287-288 (Hohl et al., 2012) and (right) the outward-facing structure of Sav1866 (Dawson and Locher, 2006) and cartoon depicting rough domain organization in the opening transition state (center). The three target positions are highlighted in spacefill on the models and as colored circles in the cartoon. CFTR domain color coding follows that of Figure 1; threonine 1246 (red), tyrosine 275 (violet), methionine 348 (blue). Blue ribbons in the homology models highlight segments 102–106 (TM1), 337–341 (TM6), 1115–1118 (TM11), and 1130–1134 (TM12), that form the narrow region of the pore (blue ovals in cartoon). Vertical colored arrow illustrates the direction and timing of the conformational wave during pore opening (early, red; late, blue).

normally dependent on correct judgement of the number of active channels in each patch.

In addition to documented reductions in unitary conductance (Linsdell, 2001), perturbations of the filter by replacement of permeating chloride with nitrate, bromide, or formate all affected gating: nitrate and bromide that bind more tightly in the pore (Linsdell, 2001) increased P_o, while formate that binds less tightly (Linsdell, 2001) decreased it (Figures 5A and 5D). Importantly, both gating effects primarily reflected changes in τ_b (Figure 5B), i.e., in rate k_{OC} of step $O_1 \rightarrow C_1$ (Figure 1B), with smaller changes in τ_{ib} (Figure 5C); the observations on nitrate replicated those of (Yeh et al., 2015). The slope of the Brønsted plot constructed from these data yielded $\Phi = 0.28 \pm 0.03$ (Figure 5E), similar to that of position 348. These ionic replacement studies therefore provide independent support for a small Φ value in the pore, confirming late movement of this region during opening.

DISCUSSION

The general structural orientations of protein domains in the stable closed and open states of the CFTR channel have been delineated by a large body of previous work. Thus, the preponderance of evidence has established that the channel's closed state corresponds to a dissociated NBD dimer interface (Vergani et al., 2005; Mense et al., 2006; Chaves and Gadsby, 2015) and inward-facing TMDs (Bai et al., 2011; Cui et al., 2014; Wang et al., 2014a), with the closed gate located on the extracellular side of the membrane (Bai et al., 2011; Cui et al., 2014; Csanády and Töröcsik, 2014; Gao and Hwang, 2015). Similarly, evidence suggests that in the open state the NBDs are dimerized (Vergani et al., 2005; Mense et al., 2006; Chaves and Gadsby, 2015), while a conducting pore is formed by outward-facing TMDs (Bai et al., 2011; Cui et al., 2014; Wang et al., 2014a). Consequently, several homology models of closed- and open-state CFTR have been constructed based on crystal structures of homologous ABC exporters in their inward- and outward-facing conformations (Mornon et al., 2009; Corradi et al., 2015) (Figure 6B, left and right). In contrast, far less is known about the nature and relative timing of the submicrosecond molecular motions that drive the channel from its closed- to its open-state conformation.

Here, we have adapted the REFER technique to obtain new insight into the dynamics of ATP-dependent gating conformational changes of the CFTR protein: careful choice of the background construct (see below) allowed selective examination of the channel opening process (step $C_1 \rightarrow O_1$, Figure 1B). The strikingly different Φ values obtained for our three target positions define a clear spatial gradient along the protein's longitudinal axis from cytoplasm to cell exterior: the very high Φ value of ~0.97 for site-2 NBD interface position 1246 (Figures 2E and Figure 6A, red) stands in stark contrast to the low Φ value of ~0.20 for intra-pore position 348. For the pore region a similarly small Φ value emerges also from our anion substitution studies: replacement of permeating chloride with anions such as nitrate and bromide, which bind more tightly to the pore (as indicated by permeability ratio measurement) (Linsdell, 2001), clearly stabilize the open state, whereas formate, which binds less tightly than chloride (Linsdell, 2001), destabilizes it (Figure 5D). Although the precise location at which permeating anions act to affect CFTR gating is unknown, this strong positive correlation between anion binding affinity in the filter and open-state stability does support the notion that the gating effects are caused by interactions of the anions with residues located somewhere in the pore. It is notable that the effects of ionic replacement on open-closed equilibrium were in each case associated with changes in closing, rather than opening rates (Figures 5B and 5C), implying that the stability of the transition state, relative to the closed state, is less sensitive to the permeating anion species. Insofar as pore-anion interactions are expected to change between the closed and open state, the implication is that in the transition state these interactions resemble those in the closed state: i.e., the pore is closed. The Brønsted plots for ionic replacement and for the 348 position closely agree with each other, and the combined data are well fitted by a single line with a slope of 0.23 ± 0.05 (Figure 6A, blue). Compared to the Φ values for the NBD1-NBD2 interface and the pore, which are close to the highest and smallest possible values for this parameter, respectively, the slope of ~0.50 of the Brønsted plot for NBD-TMD

interface position 275 (Figure 6A, violet) appears intermediate, distinctly different from the two extremes. This spatially organized Φ value gradient provides support for the interpretation that for conformational changes of folded proteins the relative magnitude of Φ reflects relative timing of ordered sequential movements (Zhou et al., 2005; Auerbach, 2007), albeit on the sub-microsecond timescale, as opposed to probabilities of taking alternative kinetic pathways known to exist for more random processes such as protein folding (Purohit et al., 2013). For the CFTR pore opening transition, this spatial Φ-gradient implies a conformational wave (Figure 6B, large vertical arrow) initiated by tightening of the NBD dimer around site 2 and propagated with some delay through the NBD-TMD interface to eventually result in pore opening.

Furthermore, this set of Φ values provides strong global constraints for the structure of the actual transition state, the highest free-energy intermediate of the channel opening process (Figure 6B, center). For the NBD interface, the Φ value of ~1 indicates that it has already reached its open-state conformation, i.e., the tight dimer is already formed (Vergani et al., 2005). In contrast, the low Φ value of ~0.2 for the pore region implies it is still in its closed-like, inward-facing conformation. Finally, the intermediate Φ value of ~0.5 for position 275 suggests that in the transition state coupling helix 2 is just on the move: it has already left its closed-like conformation but has not yet reached its open-like conformation (Figure 6B, center, bent rods). This transition state architecture, which emerges from direct measurements of relative timing of movements, confirms a previous speculation based on interpretation of enthalpy and entropy changes determined for the opening transition state (Csanády et al., 2006) but refutes the alternative proposition that during opening all structural reorganizations in the cytoplasmic loops are completed in the channel closed state (Aleksandrov et al., 2009). The large molecular strain that must arise at the NBD-TMD interface is the most likely cause of the very high enthalpy of the opening transition state ($\Delta H^{\ddagger}$ = 117 kJ/mol) and is only partially compensated by an entropy increase ($T\Delta S^{\ddagger} \geq$ 41 kJ/mol) suggested to reflect dehydration of the closing NBD dimer interface (i.e., dispersal of the layer of ordered water molecules into the disordered bulk solution) (Csanády et al., 2006). Evidently, transition-state free energy ($\Delta G^{\ddagger} = \Delta H^{\ddagger} - T\Delta S^{\ddagger}$) of wild-type CFTR is still very high, as witnessed by its very slow opening rate of ~1 s^{-1}, > 4 orders of magnitude slower than for the ligand-gated nicotinic acetylcholine receptor (Jackson et al., 1983). Moreover, it is this transient conformation of CFTR that is further destabilized (relative to the closed state) by NBD-TMD interface mutation ΔF508, causing the severe gating defect of this most common CF-associated mutant. Indeed, stabilization of the opening transition state seems an attractive strategy for designing potentiator compounds that stimulate gating of ΔF508 CFTR: thus, 5-nitro-2-(3-phenylpropylamino)benzoate, one of its most efficacious potentiators (albeit with a pore blocking side effect) (Wang et al., 2005), increases ΔF508 CFTR opening rate by precisely that mechanism (Csanády and Töröcsik, 2014).

Successful adaptation of the classical REFER approach to studying CFTR gating dynamics rested on three important innovations.

First, rather than focusing on the kinetics of pore opening and closure (Scott-Ward et al., 2007), the durations of bursts of openings and of long interburst closures were analyzed here, as the latter reflect conformational states of the pore associated with specific conformations of the NBDs: bursts are linked to tightly dimerized NBDs, while interburst closures reflect dissociation of the NBD interface around site 2 (Vergani et al., 2005; Chaves and Gadsby, 2015). The duration of the "active" pore conformation induced by a single ATP occlusion event at site 2 is also reflected by the time constant of macroscopic current relaxation upon sudden removal of ATP. Indeed, for all of our constructs that afforded macroscopic recordings, such macroscopic current relaxation time constants (Figure S2) were in good agreement with the mean burst durations obtained by conventional burst analysis of steady-state single-channel recordings (Figures 2B, 3B, 4B, and 5B), confirming that τ_b indeed reflects the duration of an activated state of the pore induced by a single ATP occlusion event.

Second, CFTR bursting follows a non-equilibrium cycle (Gunderson and Kopito, 1995; Csanády et al., 2010) (Figure 1B) to which REFER analysis is not applicable (Csanády, 2009). To study the pore opening step, we therefore employed the D1370N background mutation that truncates the gating cycle to an equilibrium scheme (Figure 1B, red frame). Indeed, this is the key feature that distinguishes our approach from previous studies and is responsible for its very different outcome. This is because in the normal hydrolytic background, mutation-induced changes in the rate of slow non-hydrolytic closure (rate k_{OC}, Figure 1B) remain unnoticed as long as the much faster hydrolytic pathway (rate $O_1 \rightarrow O_2$, Figure 1B) dominates pore closure. It is therefore not surprising that structural perturbations introduced into the nucleotide binding sites and several TMD/NBD interface positions of WT CFTR affected only channel opening rates, yielding apparent Φ values of ~1 for all positions tested (Aleksandrov et al., 2009). Similarly, previous studies identified several pore mutations that affected gating (Beck et al., 2008; Bai et al., 2010), but in the framework of a hydrolytic gating cycle even the large, almost an order of magnitude, acceleration of non-hydrolytic closing rates reported here for mutations at position 348 has so far evaded detection. Of note, the ΔF508 mutation also greatly accelerates non-hydrolytic closure (Jih et al., 2011)—suggesting an intermediate Φ value for position 508—yet under normal hydrolytic conditions ΔF508 closing rate is unaffected (Miki et al., 2010; Kopeikin et al., 2014).

Third, removal of the R domain eliminated potential confounding effects of altered R-domain-mediated gating regulation in our target-site mutants: not only does the non-phosphorylated R domain inhibit channel gating, but the phosphorylated R domain also mediates substantial stimulation of channel P_o (Winter and Welsh, 1997; Csanády et al., 2000), through mechanisms that are poorly understood. In that regard, our cut-ΔR background construct, pared down to the canonical ABC domains, reduces complexity: in addition to obviating the need for prior phosphorylation by PKA, gating of cut-ΔR CFTR is regulated only by ATP, similarly to the transport cycle of ABC exporters. Thus, our Φ-value map likely bears relevance to the transition state for the inward- to outward-facing transition in this broader family of CFTR relatives. Because gating of cut-ΔR(D1370N), like that

of WT CFTR, is strictly ATP-dependent (Figures S1 and S2), our conclusions do not necessarily apply to the mechanism of the extremely infrequent spontaneous openings observable in the absence of ATP that are promoted by certain mutations (Wang et al., 2010) and drugs (Jih and Hwang, 2013).

In conclusion, we have provided an initial characterization of the CFTR opening transition-state structure that could serve as a drug target for treating CF and developed a technique to directly measure timing of movements in distinct regions of the CFTR protein during the sub-microsecond process of channel opening. By further refining the Φ value map, this approach might be used in the future to define regions that move as a rigid body (Purohit et al., 2013, 2015), or to shed light on potentially asynchronous movements at the level of the ATP binding sites, the coupling helices, or the TM helices that form the pore.

EXPERIMENTAL PROCEDURES

pGEMHE-CFTR(837-1480(D1370N)) was constructed from pGEMHE-CFTR(837-1480), mutations at positions 275 and 348 were introduced into pGEMHE-CFTR(1-633) (Csanády et al., 2000), and mutations at position 1246 into pGEMHE-CFTR(837-1480(D1370N)) using Stratagene QuickChange. cDNA was transcribed in vitro using T7 polymerase, and 0.1–10 ng cRNA for both CFTR segments were co-injected into *Xenopus laevis* oocytes extracted from anaesthesized frogs following Institutional Animal Care Committee guidelines. Currents were recorded at 25°C in inside-out patches excised from oocytes 1–5 days after RNA injection. Pipette solution contained (in mM) 136 NMDG-Cl, 2 $MgCl_2$, 5 HEPES, pH = 7.4 with NMDG. The continuously flowing bath solution could be exchanged with a time constant <100 ms. Standard (chloride-based) bath solution contained 134 NMDG-Cl, 2 $MgCl_2$, 5 HEPES, 0.5 EGTA, pH = 7.1 with NMDG. For anion substitution experiments NMDG-Cl and $MgCl_2$ were replaced by NMDG and $Mg(OH)_2$, and the solution titrated to pH = 7.1 with nitric, hydrobromic, or formic acid, respectively. MgATP (Sigma) was added from a 400-mM aqueous stock solution (pH = 7.1 with NMDG). Unitary CFTR currents in 10 mM MgATP were recorded at −80 mV (−100 mV for formate currents) (EPC7, Heka Elektronik) at a bandwidth of 2 kHz and digitized at 10 kHz. Single-channel patches were identified as very long (typically 15 min–1 hr) recordings without superimposed channel openings. For T1246 mutants strong stimulation by 2′-deoxy-ATP at the end of each experiment was used to facilitate correct estimation of the number of active channels in the patch (Figure S4). To reconstruct bursts and interbursts, currents from patches containing no superimposed channel openings were refiltered at 20 Hz (10 Hz for anion substitution experiments), idealized by half-amplitude threshold crossing, and brief closures suppressed (Figure S3) using the method of Magleby and Pallotta (1983). Opening (k_{CO}) and closing (k_{OC}) rates were defined as $1/\tau_{ib}$ and $1/\tau_b$, respectively, and K_{eq} as k_{CO}/k_{OC}. All data are given as mean ± SEM of measurements from at least 4 (typically 5–8) long segments of single-channel recordings, from 4–13 patches for each mutant; in the face of alternating periods of lower and higher activity typical to CFTR (Bompadre et al., 2005), several hours of total recording for each construct were obtained to ensure unbiased sampling of average gating behavior. Macroscopic current ratios between 3 and 10 mM ATP were used to verify saturation by 10 mM ATP (Figure S1). Time constants of macroscopic current relaxations upon ATP removal were obtained from single-exponential fits (Figure S2).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.052.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, and software: L.C.; Resources, investigation, and formal analysis: B.S. and D.C.; Supervision and writing: L.C.

ACKNOWLEDGMENTS

We thank Beáta Töröcsik for generous help and guidance in molecular biology, and David Gadsby and Paola Vergani for critical review and discussions. Supported by MTA Lendület grant LP2012-39/2012, a Research Grant from the Cystic Fibrosis Foundation, and an International Early Career Scientist grant from the Howard Hughes Medical Institute to L.C.

Received: July 2, 2015
Revised: August 25, 2015
Accepted: September 18, 2015
Published: October 22, 2015

REFERENCES

Aleksandrov, L., Aleksandrov, A.A., Chang, X.B., and Riordan, J.R. (2002). The First Nucleotide Binding Domain of Cystic Fibrosis Transmembrane Conductance Regulator Is a Site of Stable Nucleotide Interaction, whereas the Second Is a Site of Rapid Turnover. J. Biol. Chem. *277*, 15419–15425.

Aleksandrov, A.A., Cui, L., and Riordan, J.R. (2009). Relationship between nucleotide binding and ion channel gating in cystic fibrosis transmembrane conductance regulator. J. Physiol. *587*, 2875–2886.

Auerbach, A. (2007). How to turn the reaction coordinate into time. J. Gen. Physiol. *130*, 543–546.

Bai, Y., Li, M., and Hwang, T.C. (2010). Dual roles of the sixth transmembrane segment of the CFTR chloride channel in gating and permeation. J. Gen. Physiol. *136*, 293–309.

Bai, Y., Li, M., and Hwang, T.C. (2011). Structural basis for the channel function of a degraded ABC transporter, CFTR (ABCC7). J. Gen. Physiol. *138*, 495–507.

Basso, C., Vergani, P., Nairn, A.C., and Gadsby, D.C. (2003). Prolonged nonhydrolytic interaction of nucleotide with CFTR's NH2-terminal nucleotide binding domain and its role in channel gating. J. Gen. Physiol. *122*, 333–348.

Beck, E.J., Yang, Y., Yaemsiri, S., and Raghuram, V. (2008). Conformational changes in a pore-lining helix coupled to cystic fibrosis transmembrane conductance regulator channel gating. J. Biol. Chem. *283*, 4957–4966.

Bompadre, S.G., Ai, T., Cho, J.H., Wang, X., Sohma, Y., Li, M., and Hwang, T.C. (2005). CFTR gating I: Characterization of the ATP-dependent gating of a phosphorylation-independent CFTR channel (DeltaR-CFTR). J. Gen. Physiol. *125*, 361–375.

Bozoky, Z., Krzeminski, M., Muhandiram, R., Birtley, J.R., Al-Zahrani, A., Thomas, P.J., Frizzell, R.A., Ford, R.C., and Forman-Kay, J.D. (2013). Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions. Proc. Natl. Acad. Sci. USA *110*, E4427–E4436.

Chaves, L.A.P., and Gadsby, D.C. (2015). Cysteine accessibility probes timing and extent of NBD separation along the dimer interface in gating CFTR channels. J. Gen. Physiol. *145*, 261–283.

Cheng, S.H., Gregory, R.J., Marshall, J., Paul, S., Souza, D.W., White, G.A., O'Riordan, C.R., and Smith, A.E. (1990). Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. Cell *63*, 827–834.

Corradi, V., Vergani, P., and Tieleman, D.P. (2015). Cystic fibrosis transmembrane conductance regulator (CFTR): closed and open state channel models. J. Biol. Chem. *290*, 22891–22906.

Csanády, L. (2009). Application of rate-equilibrium free energy relationship analysis to nonequilibrium ion channel gating mechanisms. J. Gen. Physiol. *134*, 129–136.

Csanády, L., and Töröcsik, B. (2014). Catalyst-like modulation of transition states for CFTR channel opening and closing: new stimulation strategy exploits nonequilibrium gating. J. Gen. Physiol. *143*, 269–287.

Csanády, L., Chan, K.W., Seto-Young, D., Kopsco, D.C., Nairn, A.C., and Gadsby, D.C. (2000). Severed channels probe regulation of gating of cystic fibrosis transmembrane conductance regulator by its cytoplasmic domains. J. Gen. Physiol. *116*, 477–500.

Csanády, L., Nairn, A.C., and Gadsby, D.C. (2006). Thermodynamics of CFTR channel gating: a spreading conformational change initiates an irreversible gating cycle. J. Gen. Physiol. *128*, 523–533.

Csanády, L., Vergani, P., and Gadsby, D.C. (2010). Strict coupling between CFTR's catalytic cycle and gating of its Cl- ion pore revealed by distributions of open channel burst durations. Proc. Natl. Acad. Sci. USA *107*, 1241–1246.

Cui, G., Rahman, K.S., Infield, D.T., Kuang, C., Prince, C.Z., and McCarty, N.A. (2014). Three charged amino acids in extracellular loop 1 are involved in maintaining the outer pore architecture of CFTR. J. Gen. Physiol. *144*, 159–179.

Dawson, R.J.P., and Locher, K.P. (2006). Structure of a bacterial multidrug ABC transporter. Nature *443*, 180–185.

El Hiani, Y., and Linsdell, P. (2010). Changes in accessibility of cytoplasmic substances to the pore associated with activation of the cystic fibrosis transmembrane conductance regulator chloride channel. J. Biol. Chem. *285*, 32126–32140.

Fatehi, M., and Linsdell, P. (2009). Novel residues lining the CFTR chloride channel pore identified by functional modification of introduced cysteines. J. Membr. Biol. *228*, 151–164.

Gao, X., and Hwang, T.C. (2015). Localizing a gate in CFTR. Proc. Natl. Acad. Sci. USA *112*, 2461–2466.

Gao, X., Bai, Y., and Hwang, T.C. (2013). Cysteine scanning of CFTR's first transmembrane segment reveals its plausible roles in gating and permeation. Biophys. J. *104*, 786–797.

Gunderson, K.L., and Kopito, R.R. (1995). Conformational states of CFTR associated with channel gating: the role ATP binding and hydrolysis. Cell *82*, 231–239.

He, L., Aleksandrov, A.A., Serohijos, A.W.R., Hegedus, T., Aleksandrov, L.A., Cui, L., Dokholyan, N.V., and Riordan, J.R. (2008). Multiple membrane-cytoplasmic domain contacts in the cystic fibrosis transmembrane conductance regulator (CFTR) mediate regulation of channel gating. J. Biol. Chem. *283*, 26383–26390.

Hohl, M., Briand, C., Grütter, M.G., and Seeger, M.A. (2012). Crystal structure of a heterodimeric ABC transporter in its inward-facing conformation. Nat. Struct. Mol. Biol. *19*, 395–402.

Hopfner, K.P., Karcher, A., Shin, D.S., Craig, L., Arthur, L.M., Carney, J.P., and Tainer, J.A. (2000). Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. Cell *101*, 789–800.

Hrycyna, C.A., Ramachandra, M., Germann, U.A., Cheng, P.W., Pastan, I., and Gottesman, M.M. (1999). Both ATP sites of human P-glycoprotein are essential but not symmetric. Biochemistry *38*, 13887–13899.

Hung, L.W., Wang, I.X., Nikaido, K., Liu, P.Q., Ames, G.F., and Kim, S.H. (1998). Crystal structure of the ATP-binding subunit of an ABC transporter. Nature *396*, 703–707.

Jackson, M.B., Wong, B.S., Morris, C.E., Lecar, H., and Christian, C.N. (1983). Successive openings of the same acetylcholine receptor channel are correlated in open time. Biophys. J. *42*, 109–114.

Jih, K.Y., and Hwang, T.C. (2013). Vx-770 potentiates CFTR function by promoting decoupling between the gating cycle and ATP hydrolysis cycle. Proc. Natl. Acad. Sci. USA *110*, 4404–4409.

Jih, K.Y., Li, M., Hwang, T.C., and Bompadre, S.G. (2011). The most common cystic fibrosis-associated mutation destabilizes the dimeric state of the nucleotide-binding domains of CFTR. J. Physiol. *589*, 2719–2731.

Kopeikin, Z., Yuksek, Z., Yang, H.Y., and Bompadre, S.G. (2014). Combined effects of VX-770 and VX-809 on several functional abnormalities of F508del-CFTR channels. J. Cyst. Fibros. *13*, 508–514.

Linsdell, P. (2001). Relationship between anion binding and anion permeability revealed by mutagenesis within the cystic fibrosis transmembrane conductance regulator chloride channel pore. J. Physiol. *531*, 51–66.

Linsdell, P., Evagelidis, A., and Hanrahan, J.W. (2000). Molecular determinants of anion selectivity in the cystic fibrosis transmembrane conductance regulator chloride channel pore. Biophys. J. *78*, 2973–2982.

Locher, K.P. (2009). Review. Structure and mechanism of ATP-binding cassette transporters. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 239–245.

Magleby, K.L., and Pallotta, B.S. (1983). Burst kinetics of single calcium-activated potassium channels in cultured rat muscle. J. Physiol. *344*, 605–623.

McCarty, N.A., and Zhang, Z.R. (2001). Identification of a region of strong discrimination in the pore of CFTR. Am. J. Physiol. *281*, L852–L867.

McDonough, S., Davidson, N., Lester, H.A., and McCarty, N.A. (1994). Novel pore-lining residues in CFTR that govern permeation and open-channel block. Neuron *13*, 623–634.

Mense, M., Vergani, P., White, D.M., Altberg, G., Nairn, A.C., and Gadsby, D.C. (2006). In vivo phosphorylation of CFTR promotes formation of a nucleotide-binding domain heterodimer. EMBO J. *25*, 4728–4739.

Miki, H., Zhou, Z., Li, M., Hwang, T.C., and Bompadre, S.G. (2010). Potentiation of disease-associated cystic fibrosis transmembrane conductance regulator mutants by hydrolyzable ATP analogs. J. Biol. Chem. *285*, 19967–19975.

Mitra, A., Cymes, G.D., and Auerbach, A. (2005). Dynamics of the acetylcholine receptor pore at the gating transition state. Proc. Natl. Acad. Sci. USA *102*, 15069–15074.

Moody, J.E., Millen, L., Binns, D., Hunt, J.F., and Thomas, P.J. (2002). Cooperative, ATP-dependent association of the nucleotide binding cassettes during the catalytic cycle of ATP-binding cassette transporters. J. Biol. Chem. *277*, 21111–21114.

Mornon, J.P., Lehn, P., and Callebaut, I. (2008). Atomic model of human cystic fibrosis transmembrane conductance regulator: membrane-spanning domains and coupling interfaces. Cell. Mol. Life Sci. *65*, 2594–2612.

Mornon, J.P., Lehn, P., and Callebaut, I. (2009). Molecular models of the open and closed states of the whole human CFTR protein. Cell. Mol. Life Sci. *66*, 3469–3486.

O'Sullivan, B.P., and Freedman, S.D. (2009). Cystic fibrosis. Lancet *373*, 1891–1904.

Purohit, P., Mitra, A., and Auerbach, A. (2007). A stepwise mechanism for acetylcholine receptor channel gating. Nature *446*, 930–933.

Purohit, P., Gupta, S., Jadey, S., and Auerbach, A. (2013). Functional anatomy of an allosteric protein. Nat. Commun. *4*, 2984.

Purohit, P., Chakraborty, S., and Auerbach, A. (2015). Function of the M1 π-helix in endplate receptor activation and desensitization. J. Physiol. *593*, 2851–2866.

Qian, F., El Hiani, Y., and Linsdell, P. (2011). Functional arrangement of the 12th transmembrane region in the CFTR chloride channel pore based on functional investigation of a cysteine-less CFTR variant. Pflugers Arch. *462*, 559–571.

Rai, V., Gaur, M., Shukla, S., Shukla, S., Ambudkar, S.V., Komath, S.S., and Prasad, R. (2006). Conserved Asp327 of walker B motif in the N-terminal nucleotide binding domain (NBD-1) of Cdr1p of Candida albicans has acquired a new role in ATP hydrolysis. Biochemistry *45*, 14726–14739.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066–1073.

Scott-Ward, T.S., Cai, Z., Dawson, E.S., Doherty, A., Da Paula, A.C., Davidson, H., Porteous, D.J., Wainwright, B.J., Amaral, M.D., Sheppard, D.N., and Boyd, A.C. (2007). Chimeric constructs endow the human CFTR Cl- channel with the gating behavior of murine CFTR. Proc. Natl. Acad. Sci. USA *104*, 16365–16370.

Smith, P.C., Karpowich, N., Millen, L., Moody, J.E., Rosen, J., Thomas, P.J., and Hunt, J.F. (2002). ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. Mol. Cell *10*, 139–149.

Tabcharani, J.A., Chang, X.B., Riordan, J.R., and Hanrahan, J.W. (1991). Phosphorylation-regulated Cl- channel in CHO cells stably expressing the cystic fibrosis gene. Nature *352*, 628–631.

Urbatsch, I.L., Beaudet, L., Carrier, I., and Gros, P. (1998). Mutations in either nucleotide-binding site of P-glycoprotein (Mdr3) prevent vanadate trapping of nucleotide at both sites. Biochemistry *37*, 4592–4602.

Vergani, P., Nairn, A.C., and Gadsby, D.C. (2003). On the mechanism of MgATP-dependent gating of CFTR Cl- channels. J. Gen. Physiol. *121*, 17–36.

Vergani, P., Lockless, S.W., Nairn, A.C., and Gadsby, D.C. (2005). CFTR channel opening by ATP-driven tight dimerization of its nucleotide-binding domains. Nature *433*, 876–880.

Wang, W., He, Z., O'Shaughnessy, T.J., Rux, J., and Reenstra, W.W. (2002). Domain-domain associations in cystic fibrosis transmembrane conductance regulator. Am. J. Physiol. Cell Physiol. *282*, C1170–C1180.

Wang, W., Li, G., Clancy, J.P., and Kirk, K.L. (2005). Activating cystic fibrosis transmembrane conductance regulator channels with pore blocker analogs. J. Biol. Chem. *280*, 23622–23630.

Wang, W., Wu, J., Bernard, K., Li, G., Wang, G., Bevensee, M.O., and Kirk, K.L. (2010). ATP-independent CFTR channel gating and allosteric modulation by phosphorylation. Proc. Natl. Acad. Sci. USA *107*, 3888–3893.

Wang, W., El Hiani, Y., and Linsdell, P. (2011). Alignment of transmembrane regions in the cystic fibrosis transmembrane conductance regulator chloride channel pore. J. Gen. Physiol. *138*, 165–178.

Wang, W., Roessler, B.C., and Kirk, K.L. (2014a). An electrostatic interaction at the tetrahelix bundle promotes phosphorylation-dependent cystic fibrosis transmembrane conductance regulator (CFTR) channel opening. J. Biol. Chem. *289*, 30364–30378.

Wang, W., El Hiani, Y., Rubaiy, H.N., and Linsdell, P. (2014b). Relative contribution of different transmembrane segments to the CFTR chloride channel pore. Pflugers Arch. *466*, 477–490.

Winter, M.C., and Welsh, M.J. (1997). Stimulation of CFTR activity by its phosphorylated R domain. Nature *389*, 294–296.

Winter, M.C., Sheppard, D.N., Carson, M.R., and Welsh, M.J. (1994). Effect of ATP concentration on CFTR Cl- channels: a kinetic analysis of channel regulation. Biophys. J. *66*, 1398–1403.

Yeh, H.I., Yeh, J.T., and Hwang, T.C. (2015). Modulation of CFTR gating by permeant ions. J. Gen. Physiol. *145*, 47–60.

Zeltwanger, S., Wang, F., Wang, G.T., Gillis, K.D., and Hwang, T.C. (1999). Gating of cystic fibrosis transmembrane conductance regulator chloride channels by adenosine triphosphate hydrolysis. Quantitative analysis of a cyclic gating scheme. J. Gen. Physiol. *113*, 541–554.

Zhang, Z.R., McDonough, S.I., and McCarty, N.A. (2000). Interaction between permeation and gating in a putative pore domain mutant in the cystic fibrosis transmembrane conductance regulator. Biophys. J. *79*, 298–313.

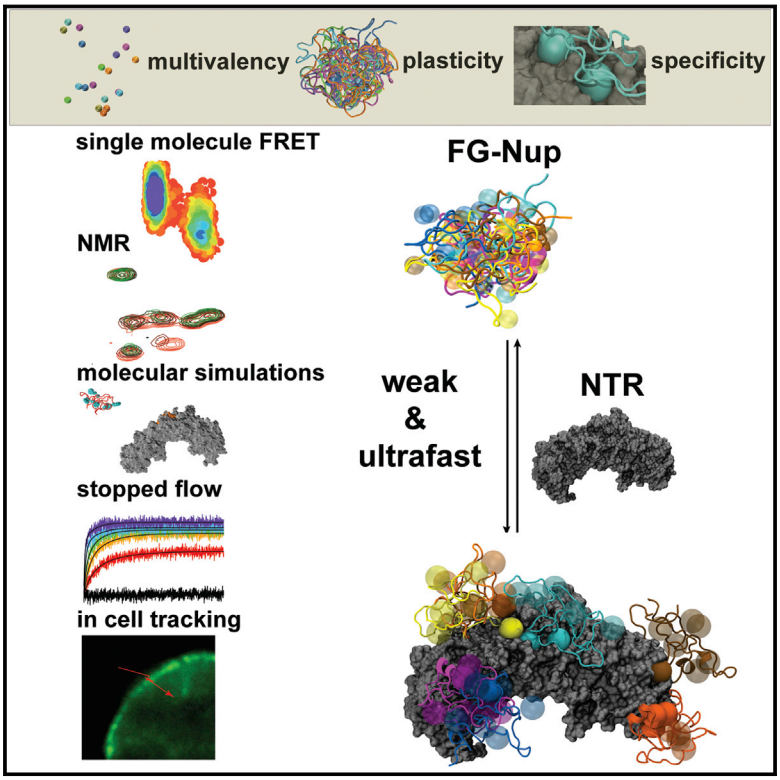
Zhou, Y., Pearson, J.E., and Auerbach, A. (2005). Phi-value analysis of a linear, sequential reaction mechanism: theory and application to ion channel gating. Biophys. J. *89*, 3680–3685.

Article

Cell

Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors

Graphical Abstract



Authors

Sigrid Milles, Davide Mercadante, Iker Valle Aramburu, ..., Martin Blackledge, Frauke Gräter, Edward A. Lemke

Correspondence

martin.blackledge@ibs.fr (M.B.), frauke.graeter@h-its.org (F.G.), lemke@embl.de (E.A.L.)

In Brief

Intrinsically disordered nucleoporins (Nups) engage rapidly with nuclear transport receptors through many minimalistic, weakly binding motifs. These Nups form polyvalent complexes while retaining conformational plasticity thus ensuring both rapid and specific transport.

Highlights

- Integrative structural biology reveals the basis of rapid nuclear transport
- Transient binding of disordered nucleoporins leaves their plasticity unaffected
- Multiple minimalistic low-affinity binding motifs create a polyvalent complex
- A highly reactive and dynamic surface permits an ultrafast binding mechanism

Milles et al., 2015, Cell *163*, 1–12
October 22, 2015
http://dx.doi.org/10.1016/j.cell.2015.09.047

CellPress

Please cite this article in press as: Milles et al., Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.047

Cell

Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors

Sigrid Milles,[1,4,5,6,8] Davide Mercadante,[2,3,8] Iker Valle Aramburu,[1,8] Malene Ringkjøbing Jensen,[4,5,6] Niccolò Banterle,[1] Christine Koehler,[1] Swati Tyagi,[1] Jane Clarke,[7] Sarah L. Shammas,[7] Martin Blackledge,[4,5,6,*] Frauke Gräter,[2,3,*] and Edward A. Lemke[1,*]

[1]Structural and Computational Biology Unit, Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]Molecular Biomechanics group, HITS gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
[3]IWR – Interdisciplinary Center for Scientific Computing, Im Neuenheimer Feld 368, 69120, Heidelberg, Germany
[4]University Grenoble Alpes, IBS, F-38044 Grenoble, France
[5]CNRS, IBS, F-38044 Grenoble, France
[6]CEA, IBS, F-38044 Grenoble, France
[7]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK
[8]Co-first author
*Correspondence: martin.blackledge@ibs.fr (M.B.), frauke.graeter@h-its.org (F.G.), lemke@embl.de (E.A.L.)
http://dx.doi.org/10.1016/j.cell.2015.09.047

SUMMARY

The mechanisms by which intrinsically disordered proteins engage in rapid and highly selective binding is a subject of considerable interest and represents a central paradigm to nuclear pore complex (NPC) function, where nuclear transport receptors (NTRs) move through the NPC by binding disordered phenylalanine-glycine-rich nucleoporins (FG-Nups). Combining single-molecule fluorescence, molecular simulations, and nuclear magnetic resonance, we show that a rapidly fluctuating FG-Nup populates an ensemble of conformations that are prone to bind NTRs with near diffusion-limited on rates, as shown by stopped-flow kinetic measurements. This is achieved using multiple, minimalistic, low-affinity binding motifs that are in rapid exchange when engaging with the NTR, allowing the FG-Nup to maintain an unexpectedly high plasticity in its bound state. We propose that these exceptional physical characteristics enable a rapid and specific transport mechanism in the physiological context, a notion supported by single molecule in-cell assays on intact NPCs.

INTRODUCTION

The plasticity of intrinsically disordered proteins (IDPs) is thought to be key to their highly diverse roles in the eukaryotic interactome and a variety of vital processes such as transcription, epigenetic regulation mechanisms, and transport through nuclear pore complexes (NPCs) (Dyson and Wright, 2005; Tompa and Fuxreiter, 2008). The central channel of the NPC is filled with phenylalanine-glycine-rich proteins, called FG-nucleoporins (FG-Nups) that are intrinsically disordered (Denning et al., 2003). FG-Nups build up an approximately 30-nm-thick permeability barrier through which large molecules (>40 kDa) can only be shuttled when bound to a nuclear transport receptor (NTR) with passage times as fast as 5 ms (Hoelz et al., 2011; Kubitscheck et al., 2005; Tu et al., 2013; Wälde and Kehlenbach, 2010). Due to the intrinsic dynamics of the FG-Nups, even state-of-the-art electron tomographic studies are not able to visualize them within the central NPC channel, despite their millimolar concentrations (Bui et al., 2013). Consequently, the molecular structure of the permeability barrier and its general mode of action are widely debated (for a review see Adams and Wente, 2013).

The key to understanding the observed nucleocytoplasmic transport phenomena resides in a description of the binding mode between FG-Nups and NTRs, for which a molecular analysis of the FG-Nup•NTR interaction is a prerequisite. Our current understanding of the molecular basis of FG-Nup•NTR interactions is in large part derived from X-ray crystallographic structures or molecular dynamics (MD) simulations of NTRs in the presence of short FG-peptides (up to ~13 amino acids in length) (Bayliss et al., 2000; Isgro and Schulten, 2005), as well as binding measurements with different NTRs or mutated NTR binding pockets (Bednenko et al., 2003; Milles and Lemke, 2014; Otsuka et al., 2008). Even for FG-Nups alone, only overall chain dimensions or long-range interactions within the Nups have so far been analyzed in solution (Milles and Lemke, 2011; Yamada et al., 2010). Notably, even such fundamental binding characteristics as the equilibrium dissociation constant (K_d) between Nups and NTRs are still matter of discussion - estimates range from a few nM to several mM (Bednenko et al., 2003; Ben-Efraim and Gerace, 2001; Tetenbaum-Novatt et al., 2012; Tu et al., 2013). However, high K_d (low affinity, ~mM) values are not easily compatible with high specificity of the transport process, while low K_d values (~nM range) cannot easily explain high transport rates, since these might be expected to correlate with long

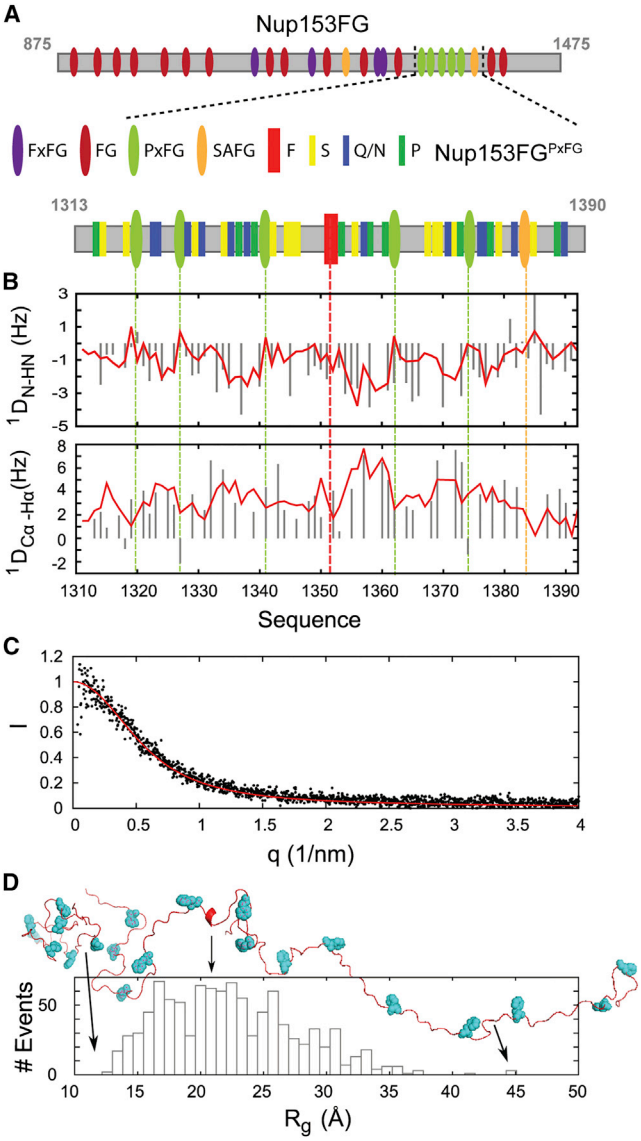


Figure 1. Conformation of Nup153FG^{PxFG}

(A) Scheme of Nup153FG constructs.

(B) Residual dipolar couplings (RDCs) of Nup153FG^{PxFG} aligned in phages. Experimentally obtained RDCs (gray bars) were compared with RDCs calculated from the ASTEROIDS ensemble obtained on the basis of experimental chemical shifts (red line). Dashed lines represent positions of FG-repeats and F1374. Color code as in (A).

(C) The same conformational ensemble was used to calculate a small angle X-ray scattering (SAXS) curve using CRYSOL (red line). The back calculated scattering curve is in good agreement with measured SAXS data under similar experimental conditions (black dots) (Mercadante et al., 2015).

(D) Distribution of the radius of gyration (R_G) from five equivalent ASTEROIDS selections. The three conformations displayed on top represent the most compact, the least compact, and one of the most prevalent conformations in the ensemble.

residence times whereas NTRs must encounter many FG-Nups while crossing the thick barrier.

Fast protein binding also typically requires proper orientation of the protein binding partners as well as conformational adaptation of the IDP to bind to a folded protein. Those can occur prior to or during binding, as described by either of the two prevalent models for protein binding namely conformational selection and induced fit (Csermely et al., 2010; Wright and Dyson, 2009). While such a conformational shift or fit can present the rate-limiting step of binding, fast binding is warranted in many biological processes. Several binding rate enhancing effects have been suggested or observed experimentally, such as maintenance of a degree of disorder (termed "fuzziness"; Tompa and Fuxreiter, 2008) by conformational funneling (Schneider et al., 2015), a large capture radius of the flexible IDPs (Shoemaker et al., 2000), and the involvement of long-range electrostatic interactions to steer (attract) proteins together (Ganguly et al., 2013).

In this work, we characterize the conformational plasticity of Nups from human and yeast in the presence of structurally and functionally diverse NTRs. A focus was a PxFG-rich domain of the Nup153 (Nup153FG^{PxFG}) as its size permitted a combination of nuclear magnetic resonance (NMR), single molecule Förster resonance energy transfer (smFRET), and molecular dynamics (MD) simulations to characterize local, residue specific, as well as long-range implications of Importinβ binding to Nup153FG^{PxFG} conformation and dynamics. Additional Brownian dynamics (BD), fluorescence stopped-flow and single molecule transport experiments with functional NPCs in permeabilized cells, revealed the detailed kinetics of the complex formation between Nup and NTR. Using this molecular, integrative structural biology approach we propose a mechanism whereby Nups contribute low-affinity minimalistic binding motifs that act in concert to create a polyvalent complex. The global Nup structure and dynamics are largely unaffected by the interaction, thereby ensuring ultrafast binding and unbinding of individual motifs—a result that explains how nuclear transport can be fast yet specific, and that may have general implications for the mechanism of action of other IDPs that exhibit a multiplicity of binding motifs.

RESULTS

Nup153FG^{PxFG} Populates a Disordered Ensemble in Solution

We initially characterized the structure and dynamics of Nup153FG^{PxFG} using high resolution NMR (Figure 1A, sequences given in Supplemental Experimental Procedure). Complete assignment of the backbone resonances (Figure S1) allowed us to develop a multi-conformational model of the protein in solution using a combination of Flexible-Meccano (Ozenne et al., 2012) and the genetic algorithm ASTEROIDS (Jensen et al., 2010). Representative ensembles comprising 200 conformers were selected on the basis of the experimental chemical shifts and were in excellent agreement with $^1D_{N-NH}$ and $^1D_{C\alpha-H\alpha}$ residual dipolar couplings and small angle X-ray scattering (SAXS) curves (Mercadante et al., 2015) that were not used in the selection process (Figures 1B–1D). The amino acid specific backbone dihedral angle distributions determined from the ensemble selections (Figure S1) show that negligible secondary structure is present.

Please cite this article in press as: Milles et al., Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors, Cell (2015), http://dx.doi.org/10.1016/j.cell.2015.09.047

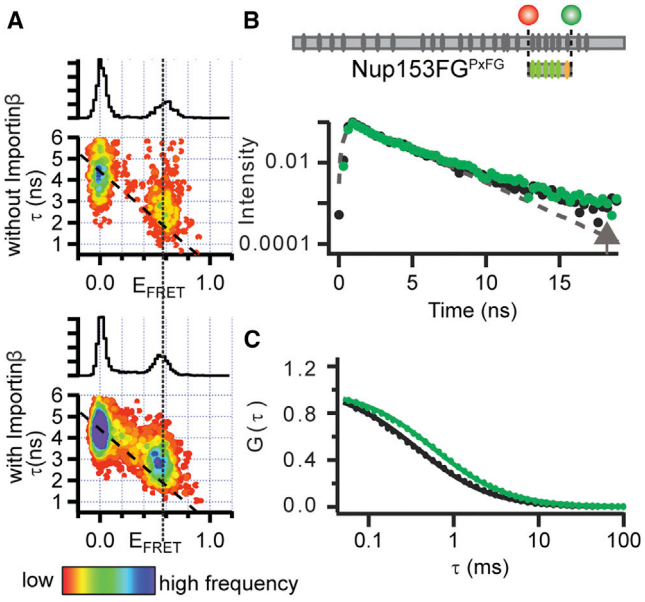


Figure 2. Nup153FG^{PxFG}·Importinβ Interaction Analyzed by smFRET

(A) FRET efficiency (E_{FRET}) versus fluorescence lifetime (τ) histograms of Nup153FG^{PxFG} in the presence and absence of Importinβ. The dotted line visualizes the center position of the FRET peak. The dashed (diagonal) lines show the static E_{FRET} relationship, on which a distribution would lie in the absence of fast dynamics.

(B) Fluorescence lifetimes (τ) of the double labeled population accumulated from single molecule data in the absence (black) and presence (green) of Importinβ. Offset from a single exponential lifetime (dashed gray curve and arrow) is a strong indicator of protein dynamics.

(C) Fluorescence correlation spectroscopy (FCS) traces retrieved from measurements of Nup153FG^{PxFG} (black dots) reflect a slower translational motion in the presence of Importinβ (green dots).

Global Structure and Dynamics of the Nup153FG^{PxFG} Are Retained upon Interaction with Importinβ as Measured by smFRET

We labeled Nup153FG^{PxFG} with a donor (Alexa488) and acceptor dye (Alexa594) for FRET at its C- and N terminus, respectively. This allowed us to measure average distance between the dyes as well as the dynamic properties of the protein using histograms relating FRET efficiency (E_{FRET}) and donor lifetimes (τ) of single molecules (sm), a method widely used to detect even minute changes in structure and dynamics, for example when IDPs bind, fold or expand (Kalinin et al., 2010; Milles and Lemke, 2011; Schuler and Eaton, 2008).

We added unlabeled Importinβ to the FRET labeled Nup153FG^{PxFG} and followed the smFRET response. While the diffusion of Nup153FG^{PxFG} in the absence and presence of Importinβ confirmed the binding of Importinβ under single molecule conditions (Figures 2 and S2), we detected neither substantial changes in E_{FRET} nor in the width of the histograms indicating absence of significant changes in the distance distribution (Figure S2 shows an all F to all A negative control). Indeed, the E_{FRET} populations of the unbound and bound Nup153FG^{PxFG} also overlay very closely with respect to τ, which indicates similarly fast dynamics of both forms (Figures 2 and S2 for detailed analysis of structure and dynamics) (Kalinin et al., 2010).

As smFRET is compatible with large proteins, we were able to repeat the same experiments for the same PxFG region within the full-length Nup153FG (601 amino acids), finding similar characteristics, and suggesting that our truncated Nup153FG^{PxFG} largely retains the conformational sampling from within the whole Nup153FG (Figure S2).

In order to determine the general nature of this binding mode, we repeated the experiments with two different FxFG-rich regions of Nup153FG, as well as the GLFG-rich yeast Nup49 and several different NTRs: i) transportin 1 (TRN1), a transport receptor involved in the import of proteins containing an M9 recognition sequence, ii) nuclear transport factor 2 (NTF2), the import receptor of RanGDP and iii) chromosomal region maintenance 1 (CRM1), a major exportin. While TRN1 and CRM1 have a similar molecular weight and superhelical structure as Importinβ, NTF2 is a much smaller, β sheet-rich dimer (Cook et al., 2007; Morrison et al., 2003). As detailed in Figure S3, despite the very distinct functionalities of the different NTRs, the smFRET and FCS measurements of the different Nups and NTRs indicate similar binding characteristics as for the Nup153FG·Importinβ complex.

Interaction with Importinβ Influences Nup153FG^{PxFG} Only Locally and Transiently

To characterize the effects of Importinβ binding on Nup153FG^{PxFG} at atomic resolution, we titrated Importinβ into a solution of ^{15}N labeled Nup153FG^{PxFG} and measured ^{1}H-^{15}N HSQC spectra at different molar ratios. Peak intensities, as well as $^{1}H^{N}$ and ^{15}N chemical shifts of Nup153FG^{PxFG}, were analyzed for each titration step (Figures 3 and S4). Resonance line broadening, associated with small changes in both $^{1}H^{N}$ and ^{15}N chemical shifts, was observed around all F's in the Nup sequence (Figure 3A). Binding was clearly highly localized, and limited to F's, with only F and the immediately adjacent amino acids being affected by the interaction. Interestingly, one single F, which is not associated with a G, is also involved in binding to Importinβ, showing the largest chemical shift changes in the ^{1}H-^{15}N HSQC spectrum during titration with Importinβ (Figure 3A and S4). ^{15}N relaxation rates measured as a function of molar ratio of Importinβ suggest that, overall, the molecule remains flexible in the complex with the transverse relaxation (R_2) increasing significantly upon Importinβ titration only around the interaction sites (Figures 3C and S4), in agreement with the above smFRET-based observations that global disorder and flexibility are not affected by Importinβ binding. Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion experiments (Figure S4) suggested that fast exchange (< 10 μs) between the bound and unbound form of Nup153FG^{PxFG} gives rise to the increased R_2 rates around the interaction sites, which makes it possible to estimate a residue-specific $K_{d,individual}$ for each position in Nup153FG^{PxFG} with Importinβ (Figures 3E, 3F and S4) from the population weighted R_2 measurements. Interestingly, the FG-specific affinities to Importinβ are not identical across the Nup153FG^{PxFG} sequence, implying a contribution of inter-FG residues to binding, although all FG-specific $K_{d,individual}$ values lie in the millimolar range.

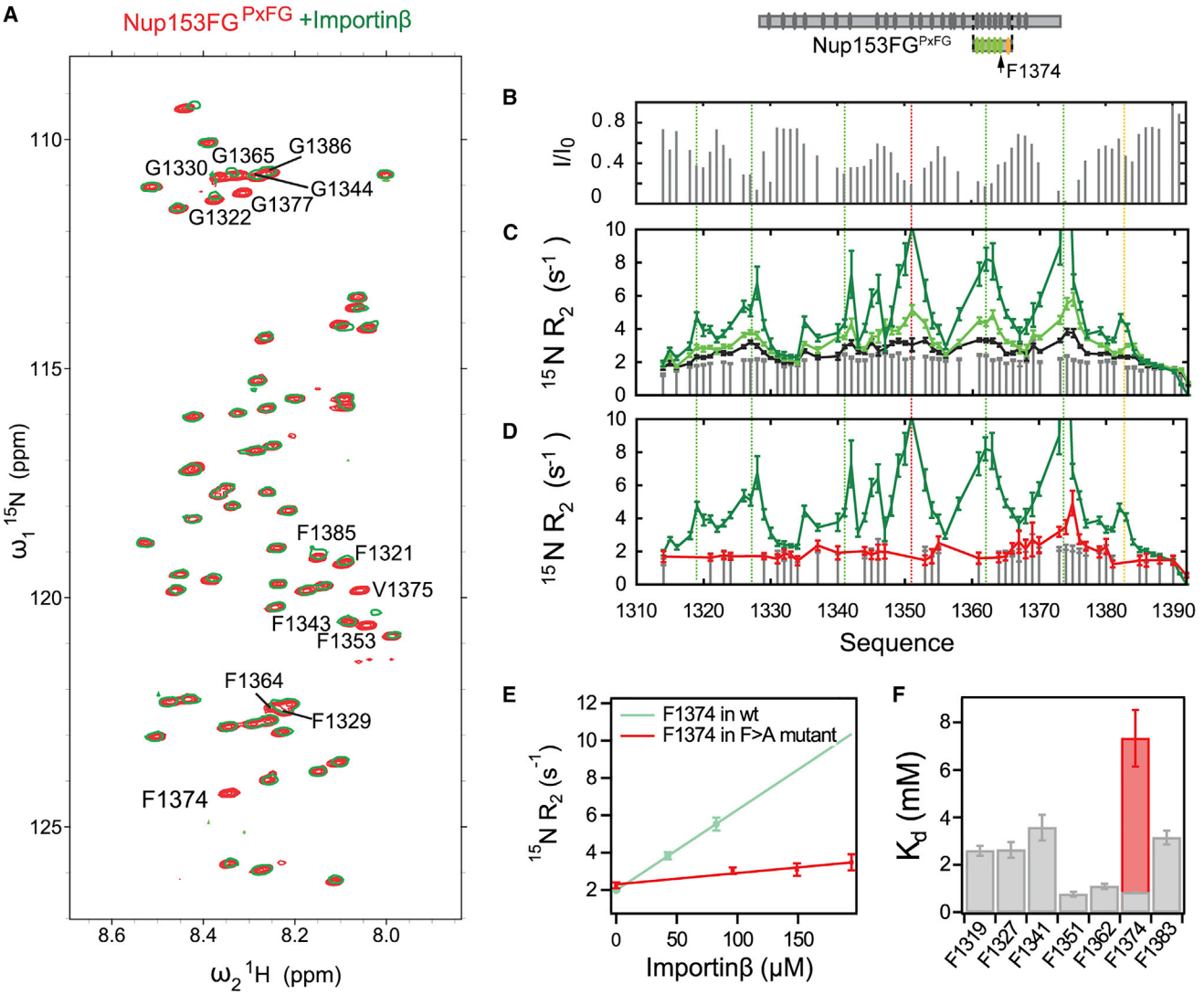


Figure 3. Nup153FG^{PxFG}·Importinβ Interaction by NMR Spectroscopy

(A) ^{1}H-^{15}N HSQC spectrum of Nup153FG^{PxFG} (red) overlayed with a spectrum of Nup153FG^{PxFG} in the presence of Importinβ (green, Nup to NTR molar ratio of 1.14, at a Nup concentration of 240 μM).

(B) The intensity ratio of the bound and unbound form of Nup153FG^{PxFG} was plotted under the same conditions as in (A).

(C) ^{15}N R_2 relaxation rates at 25°C and a ^{1}H frequency of 600 MHz were measured at different concentrations of Importinβ (gray bars are without Importinβ; black, light green and dark green at Importinβ/Nup153FG^{PxFG} molar ratios of 0.17, 0.33, and 0.72 at the constant Nup153FG^{PxFG} concentration of 250 μM).

(D) ^{15}N R_2 of Nup153AG$^{PxAG,\ F1374}$ in the absence (gray) and in the presence of Importinβ (red) overlayed with the rates for Nup153FG^{PxFG} in the presence of Importinβ under the same conditions (green).

(E) For all F in the Nup153FG^{PxFG} sequence, ^{15}N R_2 values were plotted against Importinβ concentration and fitted with a linear slope. The same analysis was performed for F1374 in Nup153AG$^{PxAG,\ F1374}$ and compared to the same F in Nup153FG^{PxFG} (compare red to green slope). R2 with errors greater than 20% were excluded from the analysis.

(F) Local K_d values were calculated from the slopes obtained in Figure S4. Gray bars correspond to K_d values obtained from Nup153FG^{PxFG}, the red bar shows the local K_d of Nup153AG$^{PxAG,\ F1374}$ binding to Importinβ.

Error bars show SD.

Strikingly, when studying the binding to different NTRs like TRN1 and NTF2 (Figure S4), despite exhibiting different binding preferences for FG-Nups (Cook et al., 2007; Milles and Lemke, 2014), their binding modes are remarkably similar to that of the Importinβ complex. The same regions in Nup153FG^{PxFG} are affected by the interaction, again with very low residue specific affinities, with the Nup remaining overall flexible when bound while interacting only locally as seen from both chemical shift changes, in the case of NTF2, and remarkably similar locally elevated transverse relaxation rates in TRN1 (Figure S4). Comparison of ^{13}C backbone chemical shifts measured in the free and NTF2-bound forms of Nup153FG^{PxFG} demonstrates that

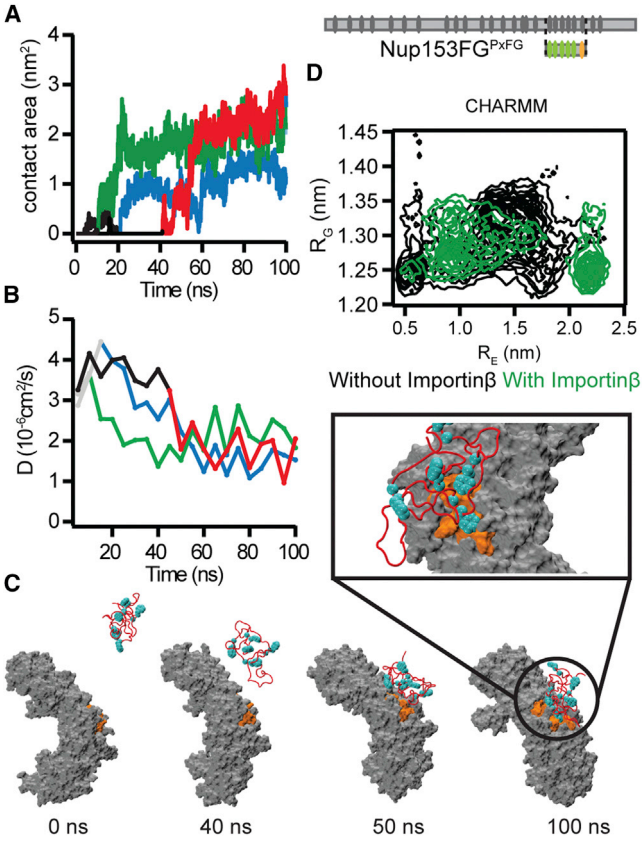


Figure 4. Binding of Nup153FG^{PxFG} to Importinβ^{N}

(A–C) Contact area between (A) Nup153FG^{PxFG} and Importinβ^{N} and (B) diffusion coefficients D as a function of time for the 4 binding events out of 10 simulations (gray/black: prior to binding; different colors: after binding; black/red curves refer to the cartoon in (C) sampled using CHARMM22* force field. (C) Snapshots collected along one of the recorded MD trajectories showing the binding between Nup153FG^{PxFG} (red cartoon) and Importinβ^{N} (gray surface). The binding sites on Importinβ^{N} and Nup153FG^{PxFG} FG-repeats are colored in orange and cyan respectively.

(D) Nup153FG^{PxFG} radius of gyration (R_G) as a function of end-to-end distance (R_E) for the unbound (black) and bound (green) ensembles of Nup153FG^{PxFG} obtained from the simulations performed using CHARMM22*.

See Figure S5 for data using the AMBER force field.

the protein backbone remains flexible upon interaction, sampling effectively the same conformational equilibrium in the free and bound state (Figure S4).

We note that during the publication process of this work, localized interaction was also reported for the yeast Nsp1 with Kap95 (the yeast homolog of Importinβ) using NMR (Hough et al., 2015), suggesting that a similar interaction mechanism may also be conserved across species.

Co-operativity of FG-Nup·Importinβ Binding

To further quantify the action of multiple FG-repeats, we designed a Nup construct, in which all F of Nup153FG^{PxFG} except F1374, the strongest interaction site for Importinβ, were replaced by A (Figure S1). Titration of Importinβ into this Nup153AG^{PxAG,F1374} mutant resulted in strongly reduced peak broadening and negligible chemical shift changes compared to Nup153FG^{PxFG} (Figure S4). As in the case of Nup153FG^{PxFG}, ^{15}N R_2 relaxation rates of Nup153AG^{PxAG,F1374} at the interaction site exhibited a linear dependence on Importinβ concentration (Figure 3E). However the effective $K_{d,individual}$ from F1374 within Nup153AG^{PxAG,F1374} reveals significantly weaker binding for this interaction site than for F1374 when situated within the wild-type (WT) protein ($K_{d,individual}$ = 7.3 mM compared to 0.8 mM, Figure 3). This result clearly shows that presenting multiple equivalent binding sites to the binding partner has a measurably positive effect on the effective affinity of the individual interaction site.

Monitoring the Nup153FG^{PxFG}·Importinβ Binding Using All-Atom MD

We employed MD simulations to investigate the experimental observations of Nup153FG^{PxFG}·Importinβ association from NMR and smFRET. From a broad ensemble of Nup153FG^{PxFG} obtained from unbiased MD simulations in explicit solvent (Movie S1), we incubated different conformers with the N-terminal portion of Importinβ (from here named Importinβ^{N} (Bayliss et al., 2000)) and monitored their binding for a total simulation time of 2 μs (Figures S5 and S6, and Table S1). The association of Nup153FG^{PxFG} to Importinβ^{N} was repeatedly observed within the simulated timescale and occurred in a specific manner (Figures 4 and S5, and Movie S2). FG-repeats docked into previously identified binding pockets on the surface of Importinβ^{N} and even formed contacts similar to those previously observed crystallographically upon interaction between Importinβ and Nsp1-derived peptides (Figures 4C and S6) (Bayliss et al., 2000). Binding was reduced and less specific for Nup153FG^{PxAG} (Figure S5), in agreement with NMR and smFRET (Figures S1, S2, and S4).

We suggest that the high solvent exposure of Fs in the unbound state (typically contained within the hydrophobic interior of folded proteins) (Figure S5) renders them readily available for Nup153FG^{PxFG}·Importinβ^{N} association, without requiring any global structural transitions in either partner (Figures 4D, S6, Movie S2).

The ability to monitor spontaneous Nup153FG^{PxFG}·Importinβ^{N} association on the sub-microsecond timescale suggests an ultrafast association (Figure S5). Underlining the generality of our observation, we were also able to monitor such a spontaneous binding event when repeating simulations for an FxFG-rich region of Nup153 binding to Importinβ^{N} (Figure S5, Movie S3, sequences given in Supplemental Experimental Procedure). However, force field inaccuracies and limited sampling prohibit the reliable extraction of an association rate, and we therefore studied the interaction further through fluorescence stopped-flow experiments (FSF) and Brownian dynamics (BD) simulations.

FSF Experiments and BD Simulations Reveal Ultrafast Binding between Nup and Importinβ

Stopped-flow kinetics monitoring fluorescence anisotropy (*r*) can be used to study binding mechanisms and measure the association rate (k_{on}) between proteins (Shammas et al., 2013). The binding of Importinβ to Nup153FG site-specifically labeled with Cy3B elicits detectable changes in *r*, due to slowed rotational

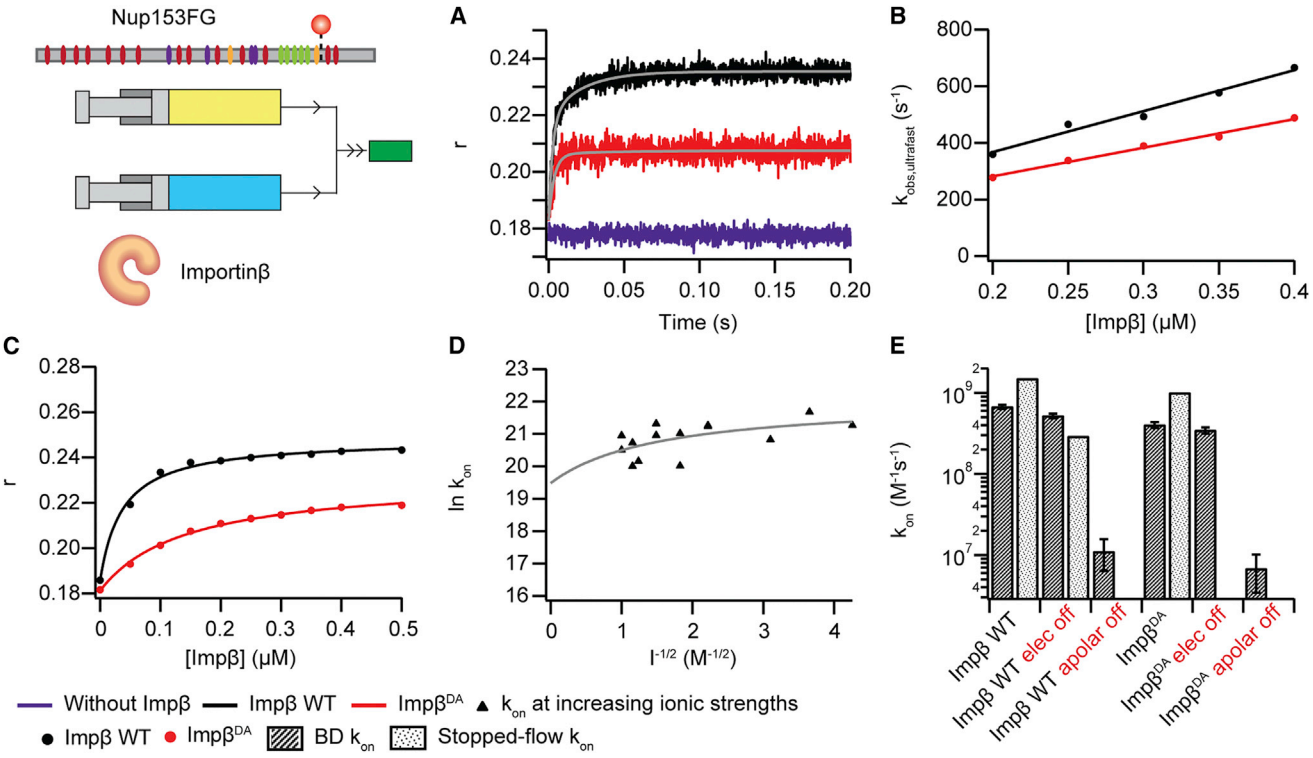


Figure 5. Association Kinetics for Nup153FG with Importinβ

(A) Stopped-flow fluorescence anisotropy was used to monitor the binding of Importinβ (Impβ) at different concentrations to Nup153FG-Cy3B. A selection of anisotropy (r) traces against time is shown for Nup153FG alone (purple) and for the binding of Importinβ WT (black) and Importinβ^{DA} (red).

(B) The observed rates ($k_{obs,ultrafast}$) from association experiments were plotted against the different Importinβ concentrations, the data were linearly fitted to obtain the association rate constants ($k_{on,ultrafast}$).

(C) Apparent $K_{d,app}$ values under the different experimental conditions.

(D) k_{on} obtained from association experiments of Nup153FG and Importinβ at different ionic strengths fitted with a Debye-Hückel-like approximation to calculate the basal rate constant at infinite ionic strength.

(E) Summary of the k_{on} values obtained from BD (dark bars) and FSF measurements (light bars) (Table S2D).

Error bars show SD.

motion (Milles and Lemke, 2014). Since Nup153FG^{PxFG} has only a very small overall binding affinity toward Importinβ, we could not detect a sufficiently strong signal change in the anisotropy measurements in the tested and experimentally feasible concentration range (Figure S7). Thus, for FSF, we used fluorescently labeled full-length Nup153FG. We performed rapid mixing experiments under pseudo-first order conditions in "physiological" transport buffer. A monoexponential function does not describe well the observed anisotropy changes in Figure 5 (Figure S7 and Table S2). This is likely a result of having multiple different binding motifs and/or the ability of multiple Importinβ to engage into binding a single Nup, which adds another level of complexity (multivalency) (Milles and Lemke, 2014; Schoch et al., 2012; Wagner et al., 2015). A biexponential equation is able to describe the kinetics, resulting in two k_{obs} per Importinβ concentration (Figures 5A, 5B, and S7). The fluorescence anisotropy at the end of the reaction was used to calculate the apparent $K_{d,app}$ (Figure 5C). Remarkably, by performing experiments at multiple NTR concentrations we extracted an ultrafast $k_{on,ultrafast}$ = $1.5 \cdot 10^9$ $M^{-1}s^{-1}$ (Figure 5B) for the major component (average amplitude of 70%), while the second component was still very fast, with a $k_{on,fast} = 6.1 \cdot 10^7$ $M^{-1}s^{-1}$ at room temperature. These FSF measurements report on overall formation of the Nup153FG• Importinβ complex i.e., one or more F binding. While we provide all results and further analysis details in Figure S7 and Table S2, for later discussion we focus on the fastest measured $k_{on,ultrafast}$.

We next estimated association rates from BD simulations, which compared to MD permit larger statistical sampling, at the cost of freezing the internal dynamics of the binding partners. Upon successful complex formation, starting from the conformations obtained from MD, the association rate was estimated (Figure S7) to be around 10^9 M^{-1} s^{-1} (Figure 5E), in agreement with stopped-flow measurements.

BD simulations carried out without the contribution of apolar desolvation generated a drastic decrease of the estimated $k_{on,BD}$ by around two orders of magnitude, while the absence of electrostatic interactions had a negligible effect (Figures 5E and S7, and Table S2D and S2E). These observations complement our evidence for an association mainly favored by the energetic gain of sequestrating F residues from the solvent and burying them into the Importinβ^N binding pockets.

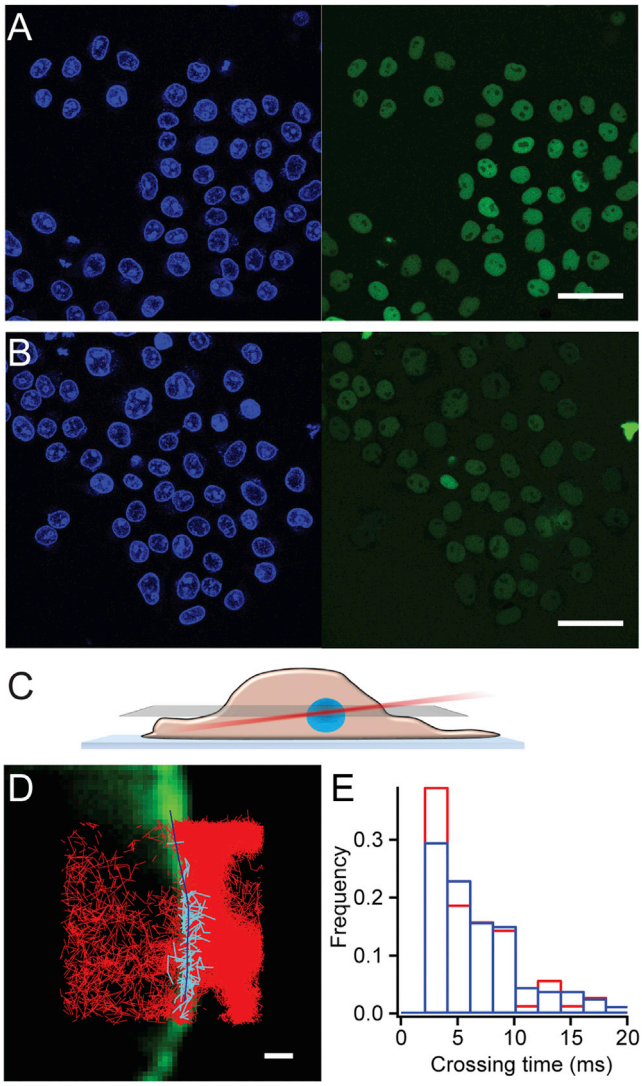


Figure 6. Nuclear Transport Assays of Importinβ and Importinβ^{DA}

(A and B) DAPI staining shown in blue, and green fluorescent cargo (NLS-MBP-eGFP) in permeabilized HeLa cells incubated with either Importinβ (A) or Importinβ^{DA} (B) (scale bar 50 μm). After 45 min, cargo accumulation is higher in the nucleus in (A).

(C) Single molecule trajectories of fluorescently labeled Importinβ were acquired in the equatorial plane of the nucleus exploiting an inclined (Hilo) illumination.

(D) Representative image of acquired single molecule trajectories of Importinβ-Alexa488 (red lines) overlaid with the ensemble image of Importinβ-Alexa647 (in green, scale bar 1μm) used to identify the nuclear envelope position (blue line). Single particle tracks of the fluorescently labeled NTR (cyan lines) crossing the nuclear envelope were analyzed to yield the characteristic barrier crossing time.

(E) The crossing time distributions reported for Importinβ (blue bars) and Importinβ^{DA} (red bars) are very fast.

While desolvation effects cannot easily be tested experimentally, high ionic strength buffers can be used to shield long-range electrostatic interactions. We thus performed a salt titration ranging from 0.05 to 1 M ionic strength (using NaCl), permitting an estimate of k_{on} under infinite electrostatic shielding by extrapolation using a Debye-Hückel-like approximation (Figures 5D and S7 and Table S2B) (Shammas et al., 2014). In line with the BD simulations, we obtained a $k_{on,elect\ off}$ of $2.9 \cdot 10^8\ M^{-1}s^{-1}$, i.e., binding remains very fast even under electrostatic shielding.

Additional stopped-flow measurements probing different Nup153FG regions (FxFG-, PxFG-rich) with diverse NTRs (NTF2, TRN1, Importinβ) are shown in Figure S7 and Table S2C. In all cases, we observed similar remarkably fast kinetics yielding consistent results for $k_{on} > 5 \cdot 10^8\ M^{-1}s^{-1}$.

Previously, solid phase binding assays indicated that the Importinβ double mutant (I178D/Y255A, termed Importinβ^{DA}) has a more than 60-fold lower K_d for binding to full-length Nup153FG as compared to Importinβ WT (Bednenko et al., 2003). $k_{on,BD}$ dropped by only 40% compared to Importinβ WT, which we confirmed by experimental FSF studies (drop of $k_{on,FSF}$ by 30%, Figure 5). However, fluorescence anisotropy measurements revealed an Importinβ^{DA} titration curve (Figure 5C) that confirms altered binding as compared to Importinβ WT, as e.g., due to an increase in k_{off}.

Single-Particle Tracking Connects Nuclear Transport of Importinβ^{DA} and Importinβ with FG-Nup Association Rates

The efficiency of an NTR to bring cargo across the NPC barrier can be assayed using standard NPC transport assays. In these assays, a fluorescent cargo (NLS-MBP-eGFP) recognized by the Importinβ transport machinery is incubated with permeabilized cells in the presence of a functional transport system and the resulting nuclear fluorescence is measured. In line with the previously reported lower K_d of Importinβ^{DA}, cargo accumulated slower compared to Importinβ WT measurements (Figures 6A and 6B) which can e.g., be due to a lower barrier crossing time, a reduced docking efficiency to the NPC or cargo release from the NPC for example.

A prediction from our kinetic analysis is that the actual speed of barrier crossing, which involves several binding and unbinding steps between NTR and FG repeats should be rather similar for WT and mutant Importinβ, as changes in k_{on} were small, and if at all, a higher k_{off} for the mutant would make crossing even faster (see discussion).

In contrast to the "bulk" transport assay, the speed of barrier crossing (characteristic crossing time) can be measured directly using single molecule (sm) tracking assays (Figure 6C), in which individual Importinβ molecules are fluorescently labeled and tracked while they cross from one side of the NPC to the other. This yielded a typical value of 6.9 ± 0.2 ms for Importinβ and 6.1 ± 0.5 ms for Importinβ^{DA} for barrier crossing (Figures 6D and 6E). We note that this crossing time is near the sampling limit of our technology, and thus faster crossing times cannot easily be captured.

DISCUSSION

The realization that many proteins are disordered has attracted considerable attention to the study of the molecular mechanisms controlling their interactions (Csermely et al., 2010; Tompa and Fuxreiter, 2008; Wright and Dyson, 2009), including the role of disorder in promoting or facilitating binding. In particular, very

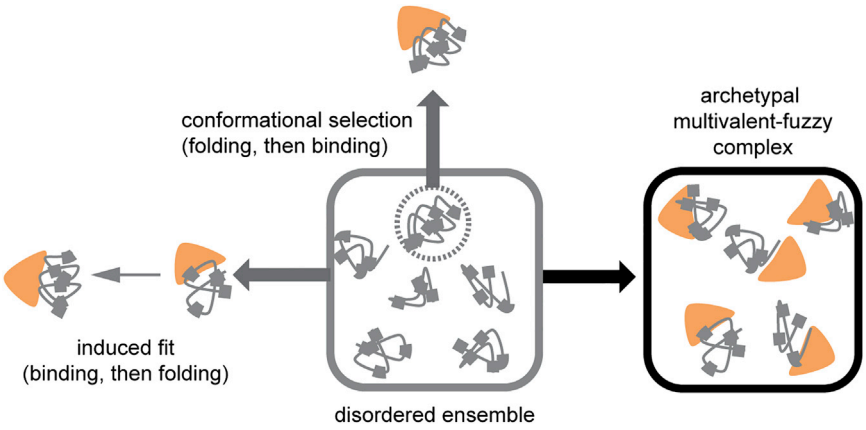


Figure 7. Binding Modes of IDPs to Folded Proteins

Schematic representation of various models describing the binding of an IDP to its folded partner. In an induced-fit mechanism the IDP partially or completely folds upon interacting with its partner, potentially showing an intermediate encounter complex as in the fly-casting mechanism (Shoemaker et al., 2000). In a conformational selection mechanism, the folded protein selects one (or several) conformation(s) of the IDP that best fits its binding pocket. These models suggest a shift in the IDP's conformational ensemble. For the Nup•NTR complex we observed formation of an "archetypal" fuzzy and multivalent complex, a binding mode that on a global scale does not require major energy or time investment for the Nup to transit from its free to the bound conformation. Note that multiple NTRs can bind one Nup and vice versa.

little is known about the binding mechanisms involved in complex processes such as nucleocytoplasmic transport, where NTRs have to engage in multiple, specific binding and unbinding events while traversing a tens of nanometer thick permeability barrier.

In this study, we have used a multidisciplinary approach to investigate the molecular mechanism underlying the interaction process between NTRs and Nups. In general, from our three core findings a coherent view emerges on how multiple rapid, yet specific protein interactions can be achieved.

Nup153FG Forms a Highly Dynamic Complex with Importinβ

Based on our smFRET measurements, we found that Nup153FG^{PxFG} resembles full-length Nup153FG with respect to its dynamics (Figures 2 and S2). Upon interaction with Importinβ, Nup153FG^{PxFG} remains flexible, engaging with Importinβ only locally, as is evident from peak broadening in the respective ^{1}H-^{15}N HSQC spectra as well as R_2 relaxation rates (Figures 3, S1, and S4). Local backbone sampling even of the interacting F was not measurably modified upon interaction. The conformers of Nup153FG^{PxFG} that were subjected to Importinβ^{N} binding in the MD simulations were also devoid of large-scale conformational changes, and interactions were only observed between individual surface exposed residues of Nup153FG^{PxFG} and Importinβ^{N}.

It appears therefore that globally, the FG-Nup maintains its conformational ensemble as shown by smFRET. This observation is sound, as IDPs frequently use motif binding to engage with their binding partners (Kragelj et al., 2015; Schneider et al., 2015; Tompa and Fuxreiter, 2008; Wright and Dyson, 2009). Our observation suggests an extraordinarily small motif (the side chain of F), which would be difficult to identify from large-scale bioinformatics approaches (Dinkel et al., 2014).

The observed binding mode appears distinct from other single motif binding interactions, as well as from mechanisms that involve global conformational transitions, such as folding upon binding (Csermely et al., 2010; Wright and Dyson, 2009) (Figure 7). The intrinsic flexibility of the Nup, the repeated occurrence and short length of the binding motif seem to create a highly reactive binding surface, which renders the individual FG-motifs prone to bind at any time without compromising the Nup's inherent plasticity.

Ultra Rapid Association of the Nup153FG·Importinβ Complex

The maximal association rate in the absence of electrostatic forces for a binary interaction system (in which all collisions are productive) can be approximated by the Einstein-Smoluchowski diffusion limit, which yields a theoretical k_{on} of $\sim 10^9$ $M^{-1}s^{-1}$ for the interaction of proteins of the size of Nup153FG and Importinβ.

Very high association rates have been observed previously in the presence of long-range electrostatic attractions (10^8-10^{10} $M^{-1}s^{-1}$) for example for the barnase/barstar interaction (Spaar et al., 2006), as well as for small IDP complexes studied by NMR (Arai et al., 2012; Schneider et al., 2015). In the absence of electrostatic steering, this upper limit is typically never reached, as successful collisions require proper orientation of the binding partners. Consequently, most experimentally observed association rates at high salt concentrations fall into the regime of 10^4–10^6 $M^{-1}s^{-1}$ (Shammas et al., 2013, 2014).

Our ensemble FSF kinetics (for Nup153FG) and BD simulations (for Nup153FG^{PxFG}) show a k_{on} of $\sim 10^9$ $M^{-1}s^{-1}$ (Figure 5) supporting the aforementioned idea of a strongly reactive binding surface. We specifically observe an influence of apolar desolvation energies in the BD simulation and electrostatics are not found to play a major role in association. This applies apparently to both, Nup153FG^{PxFG}, which is uncharged and was tested in BD, as well as Nup153FG, which has several charges in the N-terminal regions (Figures 5D and S2). Even in the limiting case of electrostatic shielding we found complex formation to still have a remarkably fast $k_{on,FSF}$ (Figures 5D, 5E and Table S2B).

While experimentally bridging the gap between our molecular-level description of the small binary Nup•NTR complex (160 kDa) in solution to the actual in vivo transport mechanisms (involving ~120 MDa NPCs) is still a challenging quest, the sm transport

experiments (Figure 6) underline that the initially unexpected kinetic findings for the Importinβ^{DA} mutant are in line with the finding in functional NPCs.

Individual FG-Repeats Bind with Low Affinity and Act in Concert for Efficient Binding

According to ensemble titration fluorescence curves, we have observed an apparent local equilibrium constant ($K_{d,app}$) between Nup153FG and Importinβ in the nanomolar regime (Figure 5C and Table S2). However, we report millimolar affinities per FG-motif from our NMR measurements within Nup153FG^{PxFG} (Figures 3 and S4), in line with a recent computational model (Tu et al., 2013). Our NMR studies further suggest that individual FG-motifs bind independently of each other, as the ^{15}N R_2 rates are similar to the values of the unbound Nup between the FG-repeats. Nevertheless, the sum of FG-motifs influences the effective binding strength of individual FGs to Importinβ, as can be seen by comparing the effective K_d for F1374 in the WT and the Nup153AG$^{PxAG,\ F1374}$ mutant (Figure 3, S1, and S4).

While these estimates of K_d values (from NMR and ensemble fluorescence) were measured on different Nup constructs, they also report on two different properties: the binding of Importinβ to a larger region of Nup153FG (fluorescence anisotropy) and to a single FG-motif (NMR), and illustrate an important characteristic of the system, namely the importance of polyvalent interactions, which is exploited also by other transport receptors (Figure S4). While an individual FG-motif might be unlikely to be bound, the chances that at least one FG-motif within the Nup molecule is bound may remain high. This stabilizing effect of multivalency/polyvalency is well known, and is even used as a design principle in enhancing the affinity of ligand interactions with multi-site targets where ligands are connected in tandem via short linkers (Brabez et al., 2011; Kramer and Karpen, 1998). Stability enhancements achieved in such experiments can approach four-to-five orders of magnitude and are primarily due to substantial decreases in the global dissociation rate, i.e., in a multivalent system the molecules only separate as a result of a dissociation event if all other motifs are unbound.

To demonstrate generality of these three core findings, we performed additional smFRET, FCS (Figures S2 and S3), NMR (Figure S4), MD (Figure S5 and Movie S3), and FSF experiments (Figure S7 and Table S2C) on a variety of different Nups from human and yeast, including the most common motif in vertebrates (FxFG) and the crucial GLFG sequence in yeast, for a diverse set of NTRs (NTF2, TRN1, CRM1, Importinβ). All results are in close agreement, highlighting the universal nature of the observed mechanism.

Currently, several models are discussed on how a permeability barrier in the NPC can be built; among those are the selective phase, the brush, the reduction of dimensionality and the karyopherin centric model, etc., as well as mixtures of those (Eisele et al., 2013; Frey and Görlich, 2007; Jovanovic-Talisman et al., 2009; Lim et al., 2007; Lowe et al., 2015; Moussavi-Baygi et al., 2011; Peters, 2009; Wagner et al., 2015; Yamada et al., 2010). These models vary mainly over how FG-Nups are arranged and potentially interlinked inside the NPC to create a tight barrier. However, common to all these models is that the concentration of FG-repeats of about 50 mM creates a very crowded environment, which is roughly in line with stoichiometric measurements of Nups and the overall size of the central channel (Bui et al., 2013; Ori et al., 2013). Independently of the transport model assumed, mobility of an NTR inside the barrier is thus largely limited by the k_{off} and k_{on} of the interaction between FG-Nups and NTRs. This is also the case if FGs interact with FGs inside the barrier as proposed in the selective phase model (Frey and Görlich, 2007), as long as these interactions are highly dynamic and do not pose a substantial energetic barrier or rate-limiting step to be melted. That we do not observe obvious FG-FG interactions in our studies is thus not necessarily inconsistent with such a model.

If we were to naively consider the characteristic time for a single Nup and Importinβ to separate based on commonly measured fast k_{on} and affinities (e.g., K_d (Nup•NTR) ~100 nM and k_{on} ~10^6 $M^{-1}s^{-1}$ → unbinding time (UT) ~100 ms), it appears impossible that Importinβ could cross a 50 mM FG-filled pore within 5 ms. This is the previously described "transport paradox," in which high specificity is somehow coupled with rapid transport (Bednenko et al., 2003; Ben-Efraim and Gerace, 2001; Tetenbaum-Novatt et al., 2012; Tu et al., 2013).

Our work (down to picosecond and atomic resolution) is largely compatible with the existing barrier models, as it addresses on a molecular mechanistic level how an NTR could rapidly pass through a dense barrier. Using a simple model of a bivalent system, we already expect an order of magnitude difference between the dissociation rate for an individual motif and that for the whole protein (Kramer and Karpen, 1998). We have also observed extremely rapid association rates (~10^9 $M^{-1}s^{-1}$) and in Supplemental Experimental Procedures (two toy models) we outline that if we consider a very rough estimate for the characteristic time for an individual motif unbinding event (UT~1 μs) for full-length Nup153 (>24 valencies), it becomes clear that the Importinβ could "creep" through the dense FG-motif plug of the pore within the short transport time. Such movement is consistent with our (Figure 6) and other NTR diffusion studies through NPCs in intact cells and various model systems (Eisele et al., 2013; Frey and Görlich, 2007; Jovanovic-Talisman et al., 2009; Moussavi-Baygi et al., 2011; Schleicher et al., 2014; Tu et al., 2013; Wagner et al., 2015).

In this case, nature has achieved a combination of high specificity with fast interaction rates. This is based on many individual low-affinity motifs paired with a binding mode that requires relatively little energy or time investment for the Nup to transit between free and bound conformations, and provides a rationale for the fast, yet specific, nuclear transport. While rapid binding can in principle be realized between proteins of single binding elements (e.g., driven by strong electrostatics), the proofreading emanating from the multiplicity and rapid repetition of many such events is what contributes to specific transport.

We note that the transport paradox goes far beyond the relevance for the transport mechanism, since transient, but targeted interactions are central to the emerging view of highly dynamic protein (and other biomolecular) interaction networks. Furthermore, FG-repeats are also present in stress and P granules (Toretsky and Wright, 2014). It seems likely that such ultrafast binding mechanisms are also important for other biological

recognition processes, where individual interaction motifs only make weak contributions, as e.g., in the recognition of glycans (Ziarek et al., 2013), or other very short linear motifs, like WG motifs in small RNA pathways (Chekulaeva et al., 2010), or binding of proteins to epigenetic marks, like many histone modifications.

In addition, ultrafast association is achieved by using the unique plasticity of multivalent disordered proteins, which is distinct from mechanisms where orientation specific binding is required for complex formation. This represents an additional biological advantage for IDPs in comparison to folded proteins, and might have further facilitated their enrichment in organisms of higher complexity.

EXPERIMENTAL PROCEDURES

Expression and Purification of Importinβ, TRN1, NTF2, CRM1 and Nup153FG

The proteins were purified essentially as described in (Milles and Lemke, 2014) following routine column chromatography and then transferred into the respective measurement buffers. Labelling of Nup153FG (amino acids 875 to 1475 of the full length Nup153; numbering with respect to the full length protein as in 'UniProt: P49790') was performed using routine procedures to introduce Alexa488 as a donor and Alexa594 as an acceptor dye for smFRET experiments (and analog for other dyes), as described in (Milles and Lemke, 2011)

NMR Studies of Nup153FG^{PxFG}

Spectral assignments of ^{13}C, ^{15}N Nup153FG^{PxFG} were obtained from a set of BEST-TROSY-type triple resonance spectra: HNCO, intra-residue HN(CA)CO, HN(CO)CA, intra-residue HNCA, HN(COCA)CB, and intra-residue HN(CA)CB (Solyom et al., 2013). For the measurements of RDCs, ^{13}C, ^{15}N Nup153FG^{PxFG} was aligned in 12 mg/ml Pf1 phages yielding a D_2O splitting of 2.16 Hz. RDCs were measured using BEST-type HNCO and HN(CO)CA experiments (Rasia et al., 2011). ^{15}N relaxation dispersion was carried out at Nup153FG^{PxFG}/Importinβ concentrations of 250 μM and 180 μM, respectively, applying CPMG frequencies between 25 and 1,000 Hz (Schneider et al., 2015). All experiments were performed in Na-phosphate buffer (pH 6), 150 mM NaCl, 2 mM DTT, 5 mM $MgCl_2$, at 25°C and at a ^1H frequency of 600 MHz if not noted otherwise.

The conformational space available to disordered Nup153FG^{PxFG} was sampled using the *Flexible-meccano* statistical coil description (Ozenne et al., 2012) and representative ensembles in agreement with experimental chemical shifts were selected using ASTEROIDS (Jensen et al., 2010) and the ensemble was subsequently cross-validated against experimental RDCs and SAXS.

SmFRET Experiments

SmFRET measurements of dual labeled freely diffusing proteins were performed on a confocal geometry detecting donor and acceptor intensities (from which the FRET efficiency E_{FRET} is calculated) as well as fluorescence lifetimes (τ) on a custom built multiparameter setup as previously described (Milles and Lemke, 2011).

Fluorescence Stopped-Flow Experiments

The association kinetics were monitored by following the fluorescence anisotropy change of Nup153FG labeled at the indicated position with Cy3B (see sequences in Supplemental Experimental Procedures) upon binding to different concentrations of NTRs, under pseudo-first order conditions. Anisotropy (r) was calculated from fluorescence intensities measured with polarizing filters in the parallel (||) and perpendicular (⊥) position.

Each trace was obtained by averaging ≥30 traces and background fluorescence was then subtracted. The anisotropy traces where fit with a biexponential function to determine k_{obs}. The different k_{obs} were plotted against the respective NTR concentrations and were linearly fit to obtain the association constant (k_{on}) from the slope.

The used BioLogic (Grenoble, France) stopped-flow equipment permits automatic titration and repeated technical replicates, which typically yield a small standard deviation. We derived an experimental error of ~20% in k_{on} measurements between different biological replicates. To be conservative, we thus do not show (the typically lower) standard deviations from technical replicates.

Transport Experiments

Routine reconstitution of the nucleocytoplasmic transport machinery in permeabilized cells was used and fluorescence cargo (NLS-MBP-eGFP) was imaged on a confocal microscope (Leica, Mannheim) at the indicated time points.

For single molecule tracking of NTRs, the same assay was used, but Importinβ-Alexa488 at single molecule concentration was tracked with an acquisition time of 2ms on a previously described home built imaging microscope (Ori et al., 2013).

All data analyses for FSF, FCS, smFRET and tracking were performed with custom written routines in IgorPro (Wavemetrics, OR).

MD and BD Simulations

The Nup153FG^{PxFG} fragment was modeled on the basis of its sequence that also included the exogenously inserted residues used for labeling of the fragment with fluorophores. For the binding simulations, Nup153FG^{PxFG} or Nup153^{FGFxFG} were randomly placed in a box of dimensions 15 × 15 × 15 nm^3 together with the N-terminal segment of Importinβ^N (PDB: 1F59). Brownian Dynamics (BD) simulations were performed starting from the MD complex that showed a specific association between the partners, and resembled the crystallographic binding pose as reported by ref. (Bayliss et al., 2000).

ACCESSION NUMBERS

The accession number for the data reported in this paper is Protein Ensemble Database (PED): 2AAE.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, two tables, and three movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.047.

AUTHOR CONTRIBUTIONS

S.M., D.M., I.V.A., designed and performed experiments, analyzed data and co-wrote the manuscript. M.R.J., N.B., C.K., S.T., J.C. provided additional reagents and analysis tools. S.L.S. designed experiments and analysis methods and co-wrote the manuscript. M.B., F.G. and E.A.L. conceived the project and co-wrote the manuscript.

ACKNOWLEDGMENTS

We are grateful for helpful comments and various discussions with Cedric Debes, Martin Beck as well as the whole Lemke group. We thank Guillaume Bouvignies for help with relaxation dispersion experiments, and Damien Maurin for sample preparation. S.M. acknowledges funding from the Boehringer Ingelheim Fonds (BIF) and an EMBO long-term fellowship (ALTF 468-2014) and EC (EMBOCOFUND2012, GA-2012-600394) via Marie Curie Action. I.V.A. acknowledges a BIF short-term fellowship. J.C. and S.L.S. are supported by the Wellcome Trust. J.C. is a Wellcome Trust Senior Research Fellow (WT/095195). E.A.L. is grateful to funds from the SFB1129 and the Emmy Noether program of the DFG, F.G. from the Klaus Tschira Foundation, and D.M. from the BIOMS program of Heidelberg University. We are also grateful to instrument access via the EMBL Pepcore facility.

Received: June 25, 2015
Revised: August 17, 2015
Accepted: September 23, 2015
Published: October 8, 2015

REFERENCES

Adams, R.L., and Wente, S.R. (2013). Uncovering nuclear pore complexity with innovation. Cell *152*, 1218–1221.

Arai, M., Ferreon, J.C., and Wright, P.E. (2012). Quantitative analysis of multisite protein-ligand interactions by NMR: binding of intrinsically disordered p53 transactivation subdomains with the TAZ2 domain of CBP. J. Am. Chem. Soc. *134*, 3792–3803.

Bayliss, R., Littlewood, T., and Stewart, M. (2000). Structural basis for the interaction between FxFG nucleoporin repeats and importin-beta in nuclear trafficking. Cell *102*, 99–108.

Bednenko, J., Cingolani, G., and Gerace, L. (2003). Importin beta contains a COOH-terminal nucleoporin binding region important for nuclear transport. J. Cell Biol. *162*, 391–401.

Ben-Efraim, I., and Gerace, L. (2001). Gradient of increasing affinity of importin beta for nucleoporins along the pathway of nuclear import. J. Cell Biol. *152*, 411–417.

Brabez, N., Lynch, R.M., Xu, L., Gillies, R.J., Chassaing, G., Lavielle, S., and Hruby, V.J. (2011). Design, synthesis, and biological studies of efficient multivalent melanotropin ligands: tools toward melanoma diagnosis and treatment. J. Med. Chem. *54*, 7375–7384.

Bui, K.H., von Appen, A., DiGuilio, A.L., Ori, A., Sparks, L., Mackmull, M.T., Bock, T., Hagen, W., Andrés-Pons, A., Glavy, J.S., and Beck, M. (2013). Integrated structural analysis of the human nuclear pore complex scaffold. Cell *155*, 1233–1243.

Chekulaeva, M., Parker, R., and Filipowicz, W. (2010). The GW/WG repeats of Drosophila GW182 function as effector motifs for miRNA-mediated repression. Nucleic Acids Res. *38*, 6673–6683.

Cook, A., Bono, F., Jinek, M., and Conti, E. (2007). Structural biology of nucleocytoplasmic transport. Annu. Rev. Biochem. *76*, 647–671.

Csermely, P., Palotai, R., and Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem. Sci. *35*, 539–546.

Denning, D.P., Patel, S.S., Uversky, V., Fink, A.L., and Rexach, M. (2003). Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. Proc. Natl. Acad. Sci. USA *100*, 2450–2455.

Dinkel, H., Van Roey, K., Michael, S., Davey, N.E., Weatheritt, R.J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., et al. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res. *42*, D259–D266.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Eisele, N.B., Labokha, A.A., Frey, S., Görlich, D., and Richter, R.P. (2013). Cohesiveness tunes assembly and morphology of FG nucleoporin domain meshworks - Implications for nuclear pore permeability. Biophys. J. *105*, 1860–1870.

Frey, S., and Görlich, D. (2007). A saturated FG-repeat hydrogel can reproduce the permeability properties of nuclear pore complexes. Cell *130*, 512–523.

Ganguly, D., Zhang, W., and Chen, J. (2013). Electrostatically accelerated encounter and folding for facile recognition of intrinsically disordered proteins. PLoS Comput. Biol. *9*, e1003363.

Hoelz, A., Debler, E.W., and Blobel, G. (2011). The structure of the nuclear pore complex. Annu. Rev. Biochem. *80*, 613–643.

Hough, L.E., Dutta, K., Sparks, S., Temel, D.B., Kamal, A., Tetenbaum-Novatt, J., Rout, M.P., and Cowburn, D. (2015). The molecular mechanism of nuclear transport revealed by atomic scale measurements. eLife *4*, 4.

Isgro, T.A., and Schulten, K. (2005). Binding dynamics of isolated nucleoporin repeat regions to importin-beta. Structure *13*, 1869–1879.

Jensen, M.R., Salmon, L., Nodet, G., and Blackledge, M. (2010). Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J. Am. Chem. Soc. *132*, 1270–1272.

Jovanovic-Talisman, T., Tetenbaum-Novatt, J., McKenney, A.S., Zilman, A., Peters, R., Rout, M.P., and Chait, B.T. (2009). Artificial nanopores that mimic the transport selectivity of the nuclear pore complex. Nature *457*, 1023–1027.

Kalinin, S., Valeri, A., Antonik, M., Felekyan, S., and Seidel, C.A. (2010). Detection of structural dynamics by FRET: a photon distribution and fluorescence lifetime analysis of systems with multiple states. J. Phys. Chem. B *114*, 7983–7995.

Kragelj, J., Palencia, A., Nanao, M.H., Maurin, D., Bouvignies, G., Blackledge, M., and Jensen, M.R. (2015). Structure and dynamics of the MKK7-JNK signaling complex. Proc. Natl. Acad. Sci. USA *112*, 3409–3414.

Kramer, R.H., and Karpen, J.W. (1998). Spanning binding sites on allosteric proteins with polymer-linked ligand dimers. Nature *395*, 710–713.

Kubitscheck, U., Grünwald, D., Hoekstra, A., Rohleder, D., Kues, T., Siebrasse, J.P., and Peters, R. (2005). Nuclear transport of single molecules: dwell times at the nuclear pore complex. J. Cell Biol. *168*, 233–243.

Lim, R.Y., Fahrenkrog, B., Köser, J., Schwarz-Herion, K., Deng, J., and Aebi, U. (2007). Nanomechanical basis of selective gating by the nuclear pore complex. Science *318*, 640–643.

Lowe, A.R., Tang, J.H., Yassif, J., Graf, M., Huang, W.Y., Groves, J.T., Weis, K., and Liphardt, J.T. (2015). Importin-β modulates the permeability of the nuclear pore complex in a Ran-dependent manner. eLife *4*, 4.

Mercadante, D., Milles, S., Fuertes, G., Svergun, D.I., Lemke, E.A., and Gräter, F. (2015). Kirkwood-Buff Approach Rescues Overcollapse of a Disordered Protein in Canonical Protein Force Fields. J. Phys. Chem. B *119*, 7975–7984.

Milles, S., and Lemke, E.A. (2011). Single molecule study of the intrinsically disordered FG-repeat nucleoporin 153. Biophys. J. *101*, 1710–1719.

Milles, S., and Lemke, E.A. (2014). Mapping multivalency and differential affinities within large intrinsically disordered protein complexes with segmental motion analysis. Angew. Chem. Int. Ed. Engl. *53*, 7364–7367.

Morrison, J., Yang, J.C., Stewart, M., and Neuhaus, D. (2003). Solution NMR study of the interaction between NTF2 and nucleoporin FxFG repeats. J. Mol. Biol. *333*, 587–603.

Moussavi-Baygi, R., Jamali, Y., Karimi, R., and Mofrad, M.R. (2011). Brownian dynamics simulation of nucleocytoplasmic transport: a coarse-grained model for the functional state of the nuclear pore complex. PLoS Comput. Biol. *7*, e1002049.

Ori, A., Banterle, N., Iskar, M., Andrés-Pons, A., Escher, C., Khanh Bui, H., Sparks, L., Solis-Mezarino, V., Rinner, O., Bork, P., et al. (2013). Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. Mol. Syst. Biol. *9*, 648.

Otsuka, S., Iwasaka, S., Yoneda, Y., Takeyasu, K., and Yoshimura, S.H. (2008). Individual binding pockets of importin-beta for FG-nucleoporins have different binding properties and different sensitivities to RanGTP. Proc. Natl. Acad. Sci. USA *105*, 16101–16106.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.R., Jensen, M.R., Segard, S., Bernadó, P., Charavay, C., and Blackledge, M. (2012). Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics *28*, 1463–1470.

Peters, R. (2009). Translocation through the nuclear pore: Kaps pave the way. BioEssays *31*, 466–477.

Rasia, R.M., Lescop, E., Palatnik, J.F., Boisbouvier, J., and Brutscher, B. (2011). Rapid measurement of residual dipolar couplings for fast fold elucidation of proteins. J. Biomol. NMR *51*, 369–378.

Schleicher, K.D., Dettmer, S.L., Kapinos, L.E., Pagliara, S., Keyser, U.F., Jeney, S., and Lim, R.Y. (2014). Selective transport control on molecular velcro made from intrinsically disordered proteins. Nat. Nanotechnol. *9*, 525–530.

Schneider, R., Maurin, D., Communie, G., Kragelj, J., Hansen, D.F., Ruigrok, R.W., Jensen, M.R., and Blackledge, M. (2015). Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. J. Am. Chem. Soc. *137*, 1220–1229.

Schoch, R.L., Kapinos, L.E., and Lim, R.Y. (2012). Nuclear transport receptor binding avidity triggers a self-healing collapse transition in FG-nucleoporin molecular brushes. Proc. Natl. Acad. Sci. USA *109*, 16911–16916.

Schuler, B., and Eaton, W.A. (2008). Protein folding studied by single-molecule FRET. Curr. Opin. Struct. Biol. *18*, 16–26.

Shammas, S.L., Travis, A.J., and Clarke, J. (2013). Remarkably fast coupled folding and binding of the intrinsically disordered transactivation domain of cMyb to CBP KIX. J. Phys. Chem. B *117*, 13346–13356.

Shammas, S.L., Travis, A.J., and Clarke, J. (2014). Allostery within a transcription coactivator is predominantly mediated through dissociation rate constants. Proc. Natl. Acad. Sci. USA *111*, 12055–12060.

Shoemaker, B.A., Portman, J.J., and Wolynes, P.G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc. Natl. Acad. Sci. USA *97*, 8868–8873.

Solyom, Z., Schwarten, M., Geist, L., Konrat, R., Willbold, D., and Brutscher, B. (2013). BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J. Biomol. NMR *55*, 311–321.

Spaar, A., Dammer, C., Gabdoulline, R.R., Wade, R.C., and Helms, V. (2006). Diffusional encounter of barnase and barstar. Biophys. J. *90*, 1913–1924.

Tetenbaum-Novatt, J., Hough, L.E., Mironska, R., McKenney, A.S., and Rout, M.P. (2012). Nucleocytoplasmic transport: a role for nonspecific competition in karyopherin-nucleoporin interactions. Mol. Cell. Proteomics *11*, 31–46.

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem. Sci. *33*, 2–8.

Toretsky, J.A., and Wright, P.E. (2014). Assemblages: functional units formed by cellular phase separation. J. Cell Biol. *206*, 579–588.

Tu, L.C., Fu, G., Zilman, A., and Musser, S.M. (2013). Large cargo transport by nuclear pores: implications for the spatial organization of FG-nucleoporins. EMBO J. *32*, 3220–3230.

Wagner, R.S., Kapinos, L.E., Marshall, N.J., Stewart, M., and Lim, R.Y. (2015). Promiscuous binding of Karyopherinβ1 modulates FG nucleoporin barrier function and expedites NTF2 transport kinetics. Biophys. J. *108*, 918–927.

Wälde, S., and Kehlenbach, R.H. (2010). The Part and the Whole: functions of nucleoporins in nucleocytoplasmic transport. Trends Cell Biol. *20*, 461–469.

Wright, P.E., and Dyson, H.J. (2009). Linking folding and binding. Curr. Opin. Struct. Biol. *19*, 31–38.

Yamada, J., Phillips, J.L., Patel, S., Goldfien, G., Calestagne-Morelli, A., Huang, H., Reza, R., Acheson, J., Krishnan, V.V., Newsam, S., et al. (2010). A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. Mol. Cell. Proteomics *9*, 2205–2224.

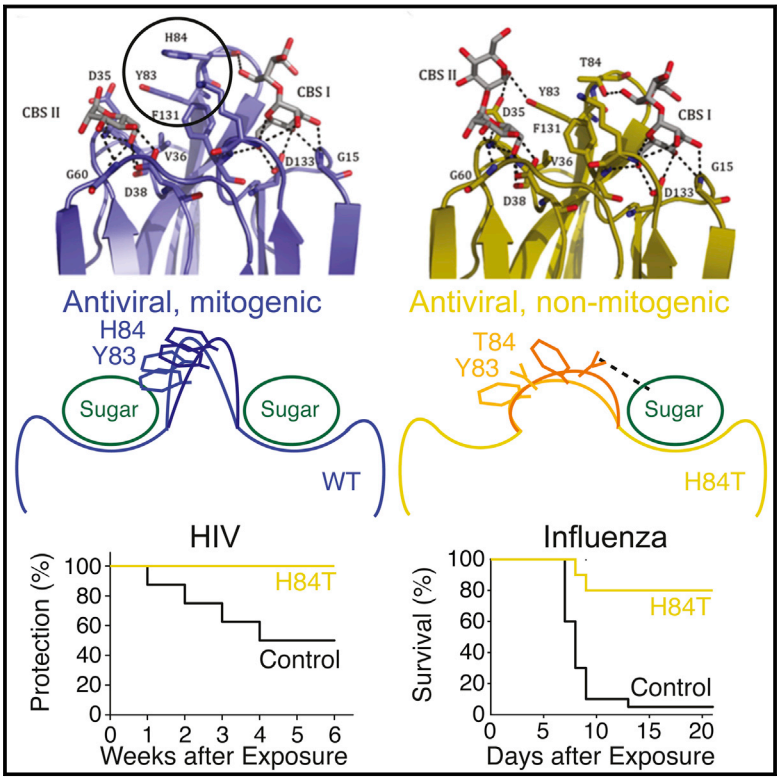
Ziarek, J.J., Veldkamp, C.T., Zhang, F., Murray, N.J., Kartz, G.A., Liang, X., Su, J., Baker, J.E., Linhardt, R.J., and Volkman, B.F. (2013). Heparin oligosaccharides inhibit chemokine (CXC motif) ligand 12 (CXCL12) cardioprotection by binding orthogonal to the dimerization interface, promoting oligomerization, and competing with the chemokine (CXC motif) receptor 4 (CXCR4) N terminus. J. Biol. Chem. *288*, 737–746.

Article

Cell

Engineering a Therapeutic Lectin by Uncoupling Mitogenicity from Antiviral Activity

Graphical Abstract



Authors

Michael D. Swanson, Daniel M. Boudreaux, Loïc Salmon, ..., Hans-Joachim Gabius, Hashim M. Al-Hashimi, David M. Markovitz

Correspondence

ha57@duke.edu (H.M.A.-H.), dmarkov@umich.edu (D.M.M.)

In Brief

Eliminating the mitogenic activity of a lectin by recalibrating how the protein "reads" surface carbohydrates expands its therapeutic potential as a broad-spectrum antiviral agent.

Highlights

- Mitogenicity and antiviral activity of a lectin can be uncoupled via mutagenesis
- The resultant lectin retains potent, broad-spectrum antiviral activity
- Pi-pi stacking of two aromatic amino acids is necessary for mitogenicity

Accession Numbers

3RFP
4PIF
4PIK
4PIT
4PIU

Swanson et al., 2015, Cell *163*, 746–758
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.056

Cell

Article

Engineering a Therapeutic Lectin by Uncoupling Mitogenicity from Antiviral Activity

Michael D. Swanson,[1,13,14] Daniel M. Boudreaux,[1,2,14] Loïc Salmon,[2,14] Jeetender Chugh,[2] Harry C. Winter,[3] Jennifer L. Meagher,[4] Sabine André,[5] Paul V. Murphy,[6] Stefan Oscarson,[7] René Roy,[8] Steven King,[1] Mark H. Kaplan,[1] Irwin J. Goldstein,[3] E. Bart Tarbet,[9] Brett L. Hurst,[9] Donald F. Smee,[9] Cynthia de la Fuente,[10] Hans-Heinrich Hoffmann,[10] Yi Xue,[11] Charles M. Rice,[10] Dominique Schols,[12] J. Victor Garcia,[13] Jeanne A. Stuckey,[4] Hans-Joachim Gabius,[5] Hashim M. Al-Hashimi,[2,11,15,*] and David M. Markovitz[1,15,*]

[1]Division of Infectious Diseases, Department of Internal Medicine, Program in Immunology, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Biophysics, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
[4]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA
[5]Institute of Physiological Chemistry, Faculty of Veterinary Medicine, Ludwig-Maximilians-University Munich, 80539 Munich, Germany
[6]School of Chemistry, National University of Ireland, Galway, Ireland
[7]Centre for Synthesis and Chemical Biology, University College Dublin, Belfield, Dublin 4, Ireland
[8]Department of Chemistry, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada
[9]Institute for Antiviral Research, Utah State University, Logan, UT 84322, USA
[10]Rockefeller University, New York, NY 10065, USA
[11]Department of Biochemistry, Duke University, Durham, NC 27710, USA
[12]Laboratory of Virology and Chemotherapy, Rega Institute for Medical Research, University of Leuven, 3000 Leuven, Belgium
[13]Division of Infectious Diseases, Department of Medicine and UNC AIDS Center, University of North Carolina, Chapel Hill, NC 27599, USA
[14]Co-first author
[15]Co-senior author
*Correspondence: ha57@duke.edu (H.M.A.-H.), dmarkov@umich.edu (D.M.M.)
http://dx.doi.org/10.1016/j.cell.2015.09.056

SUMMARY

A key effector route of the Sugar Code involves lectins that exert crucial regulatory controls by targeting distinct cellular glycans. We demonstrate that a single amino-acid substitution in a banana lectin, replacing histidine 84 with a threonine, significantly reduces its mitogenicity, while preserving its broad-spectrum antiviral potency. X-ray crystallography, NMR spectroscopy, and glycocluster assays reveal that loss of mitogenicity is strongly correlated with loss of pi-pi stacking between aromatic amino acids H84 and Y83, which removes a wall separating two carbohydrate binding sites, thus diminishing multivalent interactions. On the other hand, monovalent interactions and antiviral activity are preserved by retaining other wild-type conformational features and possibly through unique contacts involving the T84 side chain. Through such fine-tuning, target selection and downstream effects of a lectin can be modulated so as to knock down one activity, while preserving another, thus providing tools for therapeutics and for understanding the Sugar Code.

INTRODUCTION

Protein-carbohydrate interactions play essential roles in many biological processes, including adhesion and growth regulation, infection, and tumor pathogenesis (Gabius, 2015; Solís et al., 2015). Glycan-encoded information can be translated into cellular effects by receptors, termed lectins (Boyd, 1954). These carbohydrate-binding proteins are widely found in nature, have been put to considerable use in many aspects of glycobiology (André et al., 2015; Gabius et al., 2011, 2015), and have the potential to be used as antiviral agents. By specifically binding to mannosides of the glycans of glycoproteins on the surface of a virus, they can block viral attachment and/or fusion to cells.

Possible clinical applications of lectins suffer from a major drawback, the potential for side effects mediated by lectin-induced mitogenicity (Borrebaeck and Carlsson, 1989). If a mitogenic lectin were used topically in an anti-HIV microbicide, it could lead to uncomfortable inflammation, an increase in viral transmission, and even greater HIV replication because of its ability to activate T cells. Given parenterally, a mitogenic lectin could lead to systemic inflammation (Huskens et al., 2008). To date, it has remained entirely unclear whether mitogenicity and antiviral activity are dissectible in a given lectin.

We set out to rationally engineer a plant lectin isolated from the fruit of bananas (*Musa acuminata*, BanLec) (Singh et al., 2014), so as to eliminate its mitogenicity, while retaining its potent

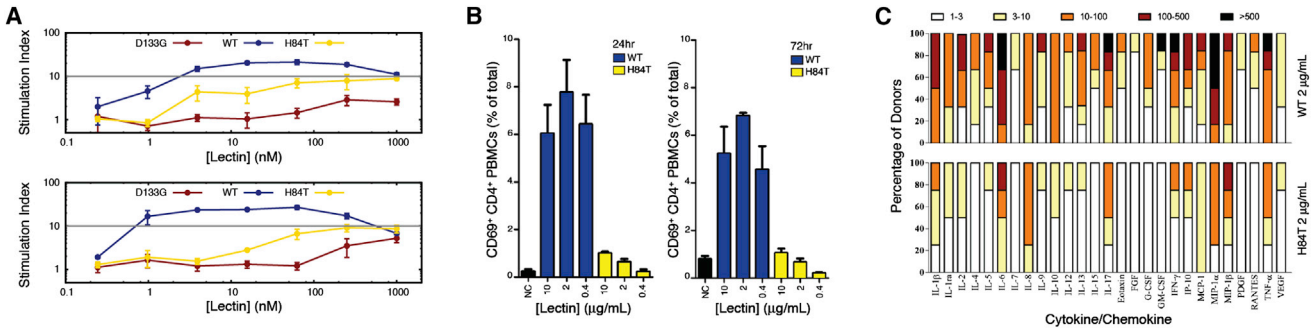


Figure 1. The H84T BanLec Mutant Is Significantly Less Mitogenic than Is WT BanLec

(A) Comparisons of the mitogenic activity of H84T to recombinant WT BanLec. PBLs from two different donors were treated with varying concentrations of lectin for 3 days and then tested for mitogenic activity by measuring BrdU incorporation by ELISA. A stimulation index of less than ten (gray line) is considered non-mitogenic. The samples for each donor were analyzed in triplicate, and error bars represent SEM. The D133G BanLec mutant, in which CBS I is altered (see Figure 4), is not mitogenic but also lacks any antiviral activity.

(B) Induction of the activation marker CD69 on CD4 T cells in the presence of WT or H84T as measured by flow cytometry, 1 or 3 days post-treatment.

(C) Induction of cytokines/chemokines by WT and H84T BanLec. PBMCs from healthy donors were incubated for 72 hr with WT or H84T BanLec at 2 μg/ml. Supernatants were collected, and cytokine levels were measured by the Bio-Plex array system. The fold-increase values of the cytokine concentrations in the supernatant of stimulated PBMCs with respect to the concentrations in the supernatant of untreated PBMCs were determined for samples from four different donors. The fold-increase values are divided into subgroups: 1- to 3-fold increase (white squares), 3- to 10-fold increase (yellow squares), 10- to 100-fold increase (orange squares), 100- to 500-fold increase (dark red squares), and >500-fold increase (black squares).

See also Figure S1.

antiviral activity. BanLec is a member of the mannose-specific jacalin-related lectin (mJRL) group that functions as a potent T cell mitogen (Singh et al., 2014). It forms a dimer with two carbohydrate-binding sites (CBS I and CBS II) in each protein subunit (Meagher et al., 2005; Singh et al., 2005). BanLec avidly associates with high-mannose-type N-glycans on the HIV-1 envelope and can thus block viral entry into cells (Swanson et al., 2010; Férir et al., 2011). Here, we show that a mutation within the sugar-binding site in BanLec makes it possible to significantly decrease mitogenic activity without compromising antiviral activity against HIV, hepatitis C virus (HCV), and influenza virus, all of which have high-mannose-type N-glycans on their surfaces. This new form of BanLec thus has the potential to be used as a broad-spectrum antiviral agent, something that is presently not available in the clinic. Further, we detail the molecular basis for separating these two distinct activities of the lectin. Our results provide proof of the feasibility of re-engineering target specificity and activity of a lectin, an approach that will greatly help to clarify how lectins read and transmit information through the Sugar Code, the biochemical platform that turns complex, sugar-encoded information into a broad spectrum of biological activities (Gabius et al., 2011; Murphy et al., 2013; Solís et al., 2015).

RESULTS

The Antiviral and Mitogenic Activities of BanLec Can Be Uncoupled through the Substitution of a Single Amino Acid

The BanLec cDNA was cloned, and the recombinant protein containing a 6x His-tag, with the sequence LEHHHHHH, expressed in *Escherichia coli*. Unless stated otherwise, all of the BanLec proteins utilized in this study are recombinant versions containing a His-tag. The recombinant His-tagged version of BanLec maintains mannose-binding properties as measured by isothermal titration calorimetry (ITC) (see discussion below) and anti-HIV activity (Figure S1). Natural BanLec is a mitogen (Gavrovic-Jankulovic et al., 2008), and we confirmed this finding with the recombinant version by exposing peripheral blood lymphocytes (PBLs) to the lectin for 3 days and measuring incorporation of bromodeoxyuridine (BrdU) (Figures 1A and S1).

To pinpoint potentially promising sites for mutational engineering, we examined crystal structures of the β-prism I structure, which is characteristic for the JRL family (Meagher et al., 2005; Singh et al., 2005). This fold consists of three Greek key structures composed of β strands; distinct loops found in the Greek keys play a role in carbohydrate binding. The first and second Greek keys include the JRL consensus motif GXXXD for sugar binding, and when mutations were introduced into these Greek keys, they abolished the mitogenic activity (as seen with the D133G mutant shown in Figure 1A), but also resulted in a loss of almost all anti-HIV activity (data not shown). The third Greek key varies among JRL members in length and sequence and is thought to play a role in binding glycan structures beyond simple saccharides (Nakamura-Tsuruta et al., 2008). H84 is part of this third loop, known to be involved in binding the second sugar moiety in α1,6-dimannosides (Singh et al., 2005). Therefore, we reasoned that altering this amino acid might result in a change in binding characteristics that would affect the lectin's mitogenic and antiviral activities differentially.

Several H84 BanLec mutants were constructed (see further discussion below), and one variant, H84T, in which the histidine is replaced by a threonine, was found to not stimulate the proliferation of lymphocytes at concentrations up to 1 μM (Figure 1A). While increased cell-surface expression of the activation marker CD69 was observed for BanLec-treated CD4+ peripheral blood mononuclear cells (PBMCs), the H84T variant induced very little

Table 1. Anti-HIV Activity Profile of BanLec, H84T BanLec, Microvirin, and the 2G12 Monoclonal Antibody in PBMCs

	HIV-1								HIV-2
	Lab Strain		Group M					Group O	
	NL4.3	BaL	B	C	F	G	H		
			US2	DJ259	BZ163	BCF-DIOUM	BCF-KITA	BCF-06	BV-5061W
	(X4)	(R5)	(R5)	(R5)	(R5)	(R5)	(R5)	(X4)	(X4)
MVN[a]	8	22	2	167	nd	nd	nd	>350	>350
BanLec[a]	0.87	0.87	1.1	2.2	2.5	6.5	3.6	14	3.7
H84T BanLec[a]	2.1	0.93	1.5	0.47	3.1	4.1	1.2	0.73	0.33
2G12 mAb[b]	140	3,710	40	>50,000	>20,000	>20,000	>20,000	>20,000	>20,000

Viral co-receptor usage (R5 or X4) is determined in U87.CD4.CCR5 and U87.CD4.CXCR4 cells and indicated in parentheses. MVN, Microvirin; nd, not determined.

[a]50% inhibitory concentration (IC_{50}) in nanomolars required to inhibit viral p24 (for HIV-1) or p27 (for HIV-2) production by 50% in PBMCs.

[b]Antibody concentration in nanograms per milliliter required to inhibit viral p24 (for HIV-1) or p27 (for HIV-2) production by 50% in PBMCs.

stimulation of this same marker (Figure 1B). Moreover, wild-type (WT) BanLec consistently caused a large increase in the induction of cytokines from the PBMCs of multiple individual donors, whereas the response to H84T was markedly reduced (Figure 1C). Thus, H84T, unlike naturally occurring and WT recombinant BanLec, is minimally mitogenic when tested by three independent methods on the peripheral blood cells of multiple different donors.

In contrast to its loss of mitogenicity, the H84T variant had an IC_{50} value against HIV in the low nanomolar range and was equally effective at inhibiting a wide range of HIV isolates as was WT BanLec, including multiple clinical isolates from different clades of group M, a group O clinical isolate, and a clinical isolate of HIV-2 (Table 1). Of note, a number of the isolates that were susceptible to H84T at low nanomolar concentrations required higher concentrations of the anti-HIV lectin Microvirin and/or were very difficult to inhibit with 2G12, a classic neutralizing anti-HIV monoclonal antibody (Table 1). Recombinant H84T without the His-tag showed very similar anti-HIV activity (data not shown).

To determine the capacity of H84T BanLec to prevent mucosal HIV transmission, we utilized the bone-marrow-liver-thymus (BLT) humanized mouse model (Wahl et al., 2012). H84T or PBS (the carrier) was topically applied to the vagina prior to challenge with HIV-1_{JR-CSF}. A total of 50% of the mice treated vaginally with PBS became infected, as determined by the presence of viral RNA in the plasma. In contrast, none of the mice treated topically with H84T showed detectable levels of viral RNA in the plasma during the course of the experiment ($p = 0.0359$; Figure 2A).

The antiviral efficacy of H84T was further evaluated against another important pathogenic virus that presents oligomannoside chains on its surface proteins, hepatitis C virus (HCV; Goffard et al., 2005). An intergenotypic HCVcc reporter virus, i.e., BiGluc-Con1/Jc1, was tested in Huh-7.5 cells (Figure S2) (Reyes-del Valle et al., 2012). The addition of H84T to the inoculum decreased HCV in a dose-dependent manner and to levels comparable to inhibition by CD81 antibody, a positive control that blocks the cellular receptor for HCV (Figure S2A; data not shown). Co-incubation of virus inoculum with the BanLec derivative D133G/38A, which, similar to the D133G mutant (Figure 1A), is inactive, was found to not decrease viral replication (Figure S2B). At the EC_{90} concentration (determined in Huh-7.5 cells), H84T also reduced HCV replication to levels similar to neutralizing E2 antibody in a primary human fetal liver culture (data not shown). Finally, to determine if the H84T-specific reduction of HCV was due to inhibition of viral RNA replication, the effect of H84T BanLec was monitored in Huh-7.5 CD81 knockdown cells ($CD81^{lo}$; Figure S2C). In this single-cycle assay, H84T decreased HCV replication over time in the control cell background only, further supporting the hypothesis that H84T inhibits viral replication at entry (receptor binding, membrane fusion), consistent with what we previously observed with WT BanLec against HIV (Swanson et al., 2010).

Glycosylation sites on the HCV E1 and E2 envelope proteins are highly conserved across genotypes (Goffard et al., 2005). Utilizing a panel of chimeric *Gaussia* luciferase reporter viruses, in which the structural region (core-NS2) was encoded by differing genotypes, H84T was observed to decrease HCV replication in a dose-dependent manner (Figures 2B–2J; Table S1). H84T BanLec thus appears to be a pan-genotypic inhibitor of HCV infection.

The hemagglutinin of influenza A viruses bears high-mannose-type N-glycans that are susceptible to host lectins (Ng et al., 2012). In studies employing a retroviral core pseudotyped with the hemagglutinins of the 1918 H1N1 and the H5N1 avian pandemic influenza viruses, WT and H84T BanLec were both very active and equally inhibitory (Figures 2K and 2L).

Next, we found that H84T BanLec is very active against multiple WT strains of influenza A tested in MDCK cells in tissue culture. Significant activity was seen against A/California/04/2009 (H1N1 pandemic strain), California/07/2009 (H1N1 pandemic strain), A/New York/18/2009 (H1N1 pandemic strain), and Perth/16/2009 (H3N2) with EC_{50} values of 1–4 μg/ml versus H1N1 virus and 0.06–0.1 versus H3N2 virus. A mutant form of BanLec that does not bind mannose, D133G/D38A, had no activity, excluding carbohydrate-independent effects. Importantly, significant activity was also seen with H84T against the Duck/MN/1525/81 H5N1 avian strain (EC_{50} of 5–11 μg/ml), confirming our results obtained with pseudotyped virus (Figure 2L). Finally, as some mouse-adapted strains of influenza lack mannose on their hemagglutinin (Smee et al., 2008), we tested an H1N1

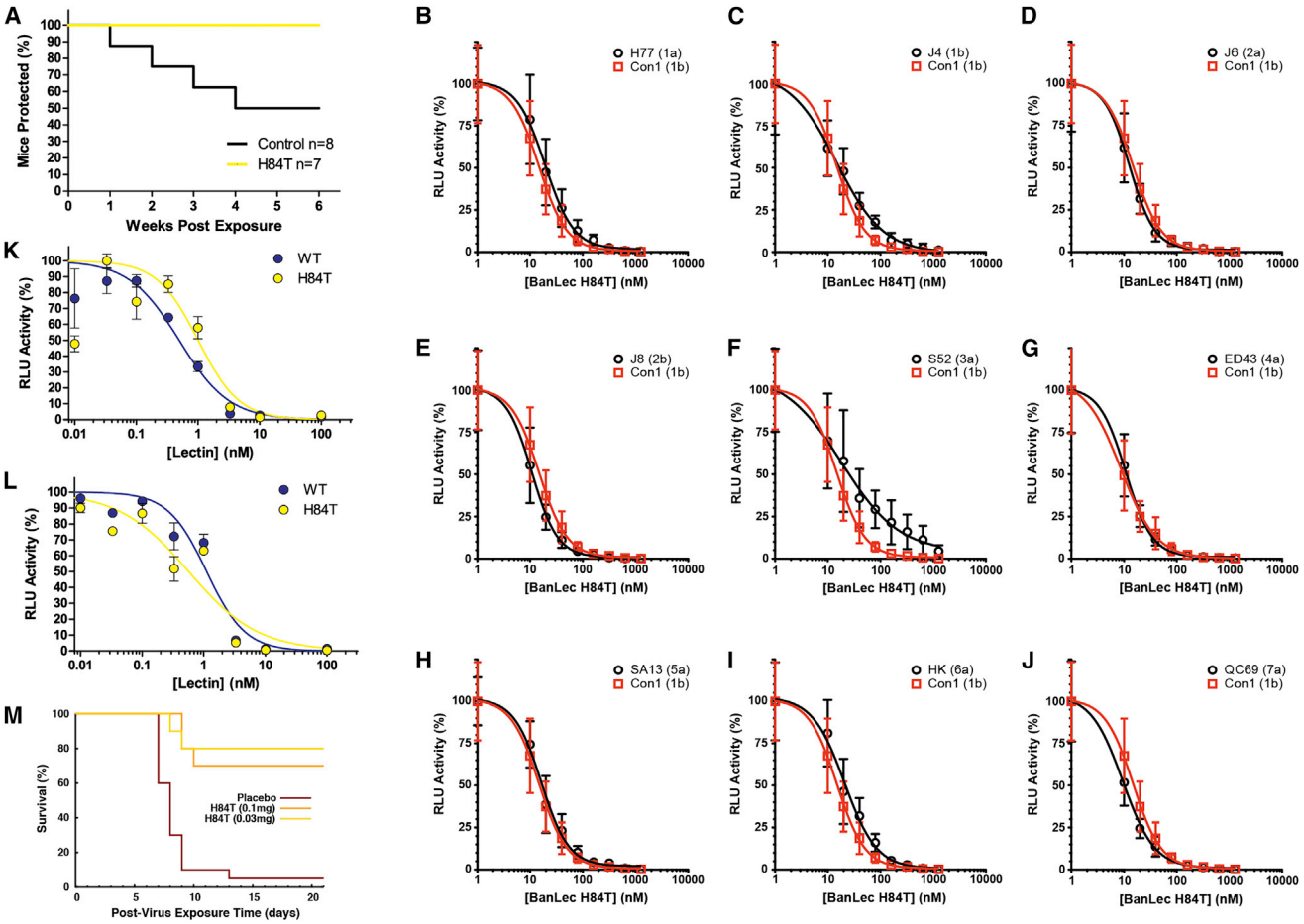


Figure 2. H84T BanLec Has Potent Antiviral Activity In Vitro and In Vivo

(A) Protection from vaginal HIV-1JR-CSF infection of BLT humanized mice by H84T BanLec. Mice were vaginally exposed to HIV in the presence or absence of topical H84T. HIV infection was determined by the presence of plasma viral load over a period of observation of 6 weeks. The times to plasma viremia were then combined to generate a Kaplan-Meier plot of the protection from vaginal HIV infection provided by H84T BanLec. Log rank analysis ($p = 0.0359$) confirmed that topically administered H84T prevents vaginal HIV-1 JR-CSF infection in BLT mice.

(B–J) Increasing concentrations of H84T (0, 10, 20, 40, 80, 160, 320, 640, and 1,280 nM) were mixed with the indicated HCVcc inoculum at a MOI of 0.1 or 0.05. After 6 hr incubation, cells were washed and media containing additional lectin was added. At 72 hr post-infection, HCV replication was analyzed by luciferase activity in supernatants. All HCVcc were bicistronic *Gaussia* luciferase reporter genomes, of which structural proteins were encoded by differing genotypes as indicated. The means and SD are plotted for two independent experiments containing five replicates each. The corresponding $EC_{50/90}$ values and their respective confidence intervals were determined and are displayed in Table S1. See also Figure S2.

(K) The activity of WT or H84T BanLec against the 1918 H1N1 pandemic influenza strain as measured by luciferase assay in the pseudotyped virus system described in the Experimental Procedures.

(L) The activity of WT and H84T against the H5N1 avian influenza strain as assessed in (K).

(M). Survival of mice challenged intranasally with influenza and then treated with H84T BanLec or control intranasally 4 hr after challenge and then daily for 5 days.

(A/WSN/1933) isolate previously shown to be inhibited by mannose-binding proteins for its sensitivity to our new agent. H84T was indeed quite active against this H1N1 strain, which causes disease in mice. Most importantly, we found that intranasal (IN) H84T BanLec, first given 4 hr after IN viral challenge, effectively blocks influenza infection in the mouse model (Figure 2M). Taken together, studies with pseudotyped virus, WT virus in tissue culture, and a mouse model of influenza demonstrate significant activity of H84T against multiple strains of influenza.

H84T BanLec Is Less Active in Multivalent Interactions

To begin to delineate the basis for the H84T mutant protein's markedly decreased mitogenic and pro-inflammatory activity, while yet maintaining its potent antiviral capacity, binding properties of H84T and WT BanLec to monovalent sugars in solution were compared. The association constants (K_a) measured using isothermal titration calorimetry (ITC) for binding to methyl α-D-mannopyranoside were similar for recombinant His-tagged WT (383 mM^{-1}) and H84T (353 mM^{-1}) and were consistent with previous measurements for naturally occurring BanLec

(333 mM^{-1}) (Mo et al., 2001; Winter et al., 2005). Interestingly, slightly weaker affinities were observed for H84T as compared to WT when analyzing binding to dimannoside (300 versus 227 mM^{-1} for WT and H84T, respectively) (Table S2).

As mitogenicity involves cross-linking of distinct counterreceptors on cell surfaces that trigger outside-in signaling, the loss of mitogenicity seen with H84T and its slightly diminished binding affinity for disaccharides compared to monosaccharides suggested that the biological differences between the two proteins might arise due to the differences in their binding properties to more complex glycans. A simple assay that provides insight into binding to cell-surface glycans and cross-linking activity (here in *trans*, that is, between cells) is measuring lectin-induced aggregate formation of erythrocytes. The minimal concentrations for agglutination were found to be significantly different, i.e., at 3 and 437 μg/ml for WT and H84T, respectively (Table S2). This result reveals a marked disparity in building stable aggregates based on more than monovalent interactions with cell-surface mannosides.

Synthetic glycoclusters are excellent tools that range in size from bivalent compounds to glycodendrimersomes (Murphy et al., 2013; Solís et al., 2015), so their locally increased density of ligands will trace a change in the interaction/association profile when testing WT and variant proteins under identical conditions. The association of a lectin with a ligand-bearing surface is sensitive to the presence of haptenic sugar, and its presentation in local clusters can enhance its inhibitory capacity. Mimicking the natural display of high-affinity ligands, synthetic glycoconjugates (carbohydrates attached to a scaffold enabling oligo- to polyvalency) thus are able to interfere with lectin binding to ligand-presenting surfaces in quantitative terms. The design of glycoclusters and the determination of their inhibitory activity on lectin binding (to glycoproteins or to a cell), measured as the inhibitory concentration (IC) at which the extent of lectin binding to a glycoligand is reduced by 50% (IC_{50} value), provide a measure of the avidity of a lectin for multivalent associations. In total, we tested a panel of 11 bi- to dodecavalent glycoclusters systematically in titrations in two types of assay, one biochemical and one cellular. In both cases, the mannose-specific lectin concanavalin A was used as positive control, and lectin binding to the glycan-presenting matrix was ascertained to be saturable and dependent on carbohydrate presence.

First, we established a surface rich in presentation of mannose residues. A neoglycoprotein (a conjugate of albumin and mannose derivatives) was adsorbed to the plastic surface of microtiter plate wells, building the matrix for letting the biotinylated lectins dock. The surface-associated label was then quantitatively assessed spectrophotometrically. Titrations of the extent of binding with increasing amounts of inhibitor were performed to determine the IC_{50} value; the glycoclusters (Figure 3A) were individually tested. As demonstrated by the example shown in Figure 3B, these experiments allowed us to determine IC_{50} values as a measure for sensitivity of lectin binding in the presence of inhibitors. Binding of the H84T mutant was found to be much more susceptible to glycocluster inhibition than was surface contact formation of the WT BanLec, consistent with the lower cross-linking capacity in hemagglutination (Tables S2 and S3).

To confirm the above and increase the biological relevance of the findings, we proceeded to monitor cell binding, using the surface of cultured cells as a platform for contact of the labeled lectins. Tested under identical conditions, WT reacted more strongly with cells than did H84T (Figures 3Ca and 3Cb). In addition to testing the physiologic glycome profile on the cells, we increased the level of lectin-reactive high-mannose-type N-glycans by treating the cells with the α-mannosidase I inhibitor 1-deoxymannojirimycin. Enhanced binding of both proteins was seen (Figures 3Cc and 3Cd), with the difference in mean fluorescence intensity between H84T and WT being maintained. Thus, increased ligand availability did not reduce the relative difference between H84T and WT proteins. Glycocluster testing on cells, for example, the tetravalent compound **11** (Figures 3Ce and 3Cf), fully confirmed the differential sensitivity seen in the solid-phase assays. These results are completely consistent with the decreased capacity for H84T BanLec to agglutinate erythrocytes and further confirm that H84T and WT differentially interact with multivalent surfaces, but not with the monosaccharide.

High-Resolution X-Ray Structures Reveal Loss of Pi-Pi Stacking between Y83 and H84 and an Altered Sugar Contact Profile in H84T

To examine the structural basis for the difference in carbohydrate-binding modes between WT and H84T, we determined the crystal structures of the recombinant proteins both in the absence and in the presence of dimannoside (M2) (Figure 4). The X-ray structure of recombinant WT BanLec was very similar to its naturally occurring, purified counterpart (Meagher et al., 2005), consistent with the similar biological activities of the two proteins. The monomer forms a β-prism I fold containing three Greek key motifs with 3-fold symmetry and two carbohydrate-binding sites (CBS I and II). CBS I consists of loops on the top of the first Greek key; CBS II sits on the top of the second Greek key. The two binding sites are separated by a loop (residues 83–88) within the third Greek key (Figure 4A), which has been suggested to be an important determinant of carbohydrate binding specificity (Meagher et al., 2005); H84 is within this loop. It is worth noting that glycerol units were observed in the different binding sites of the WT protein.

Both recombinant His-tagged proteins (WT and H84T) and the WT from bananas form a dimeric structure with interface between β strand 1 (residues 4–10), β strand 10 (residues 110–118), and two C-terminal residues (E140 and P141) from each monomer, resulting in a quasi-eight-stranded β sandwich structure. The presence of the C-terminal His-tag on recombinant WT and H84T neither altered the dimer interface nor did its presence disrupt the non-biological asymmetric tetramer that formed due to crystal packing in all the reported BanLec crystals.

Apo WT and H84T form very similar structures as indicated by an overall root-mean-square deviation of 0.26 Å. Nevertheless, there are significant differences in and around the site of mutation. In WT, H84 stacks on Y83 to form a pi-pi stacking interaction that directs both residues toward CBS II, resulting in a "wall" that separates the two CBS (Figure 4B). In sharp contrast, in H84T, no pi-pi stacking can occur (Figure 4C). Instead, the threonine side chain points toward CBS I (Figure 4C). In WT,

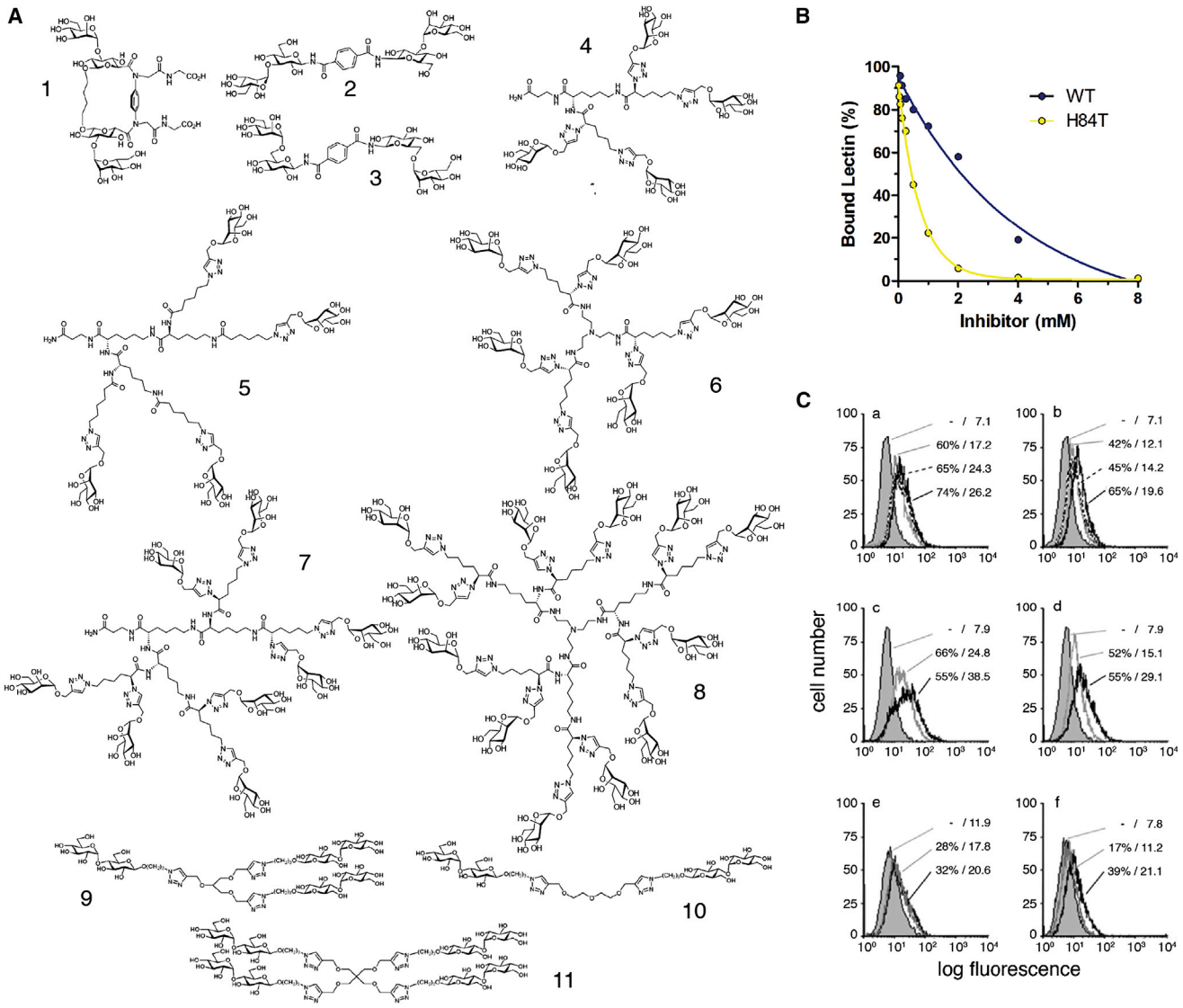


Figure 3. Binding of H84T and WT BanLec to Glycoclusters

(A) Structures of the tested glycoclusters.

(B) Titration curves for relative signal intensity, reflecting the extent of binding of WT (blue) and H84T mutant (yellow) BanLec proteins to surface-immobilized neoglycoprotein in the presence of increasing amounts of the tetravalent maltose-presenting glycocluster (**11**).

(C) Semilogarithmic illustration of fluorescent surface staining of human SW480 colon adenocarcinoma cells by labeled WT (left) or H84T (right) BanLec. Control for background (0% value) is given as the gray area, and quantitative data (percentage of positive cells/mean fluorescence intensity) are presented. Lectin staining was monitored with increasing concentrations (1, 2, and 5 μg/ml; given in a and b), at 2 μg/ml with cells without (gray) or after treatment (black) with 1-deoxymannojirimycin (c and d), and at 1.5 (WT) or 3 (H84T) μg/ml with the tetravalent glycocluster **11** at 1 mM (WT) or 0.75 mM (H84T) (e and f).

See also Tables S2 and S3.

the H84/Y83 stack prevents the side chain of residue 84 from pointing toward the CBS I.

The X-ray structures of WT and H84T bound to a dimannoside (M2) feature two dimers in the asymmetric unit forming a non-biological asymmetric tetramer, and four sets of CBS each bound to a dimannoside molecule. The position of the first mannose moiety of M2 is well resolved in the electron density maps of CBS I and II of both proteins, suggesting that it is tightly bound to both structures (Figures 4B, 4C, and S3). In CBS I of the WT and H84T, there are five hydrogen bonds (H-bonds) between each protein and the first mannose moiety, involving OD1 and OD2 of D133 and the backbone N of G15, K130, and F131. In CBS II, there are six H-bonds stabilizing the position of the saccharide, which include side-chain atoms OD1 and OD2 of D38 and the backbone N of N35, V36, and G60.

The main difference in ligand binding between the proteins involves the second mannose moiety that is more accessible to solvent and residue 84. This second mannose moiety gives

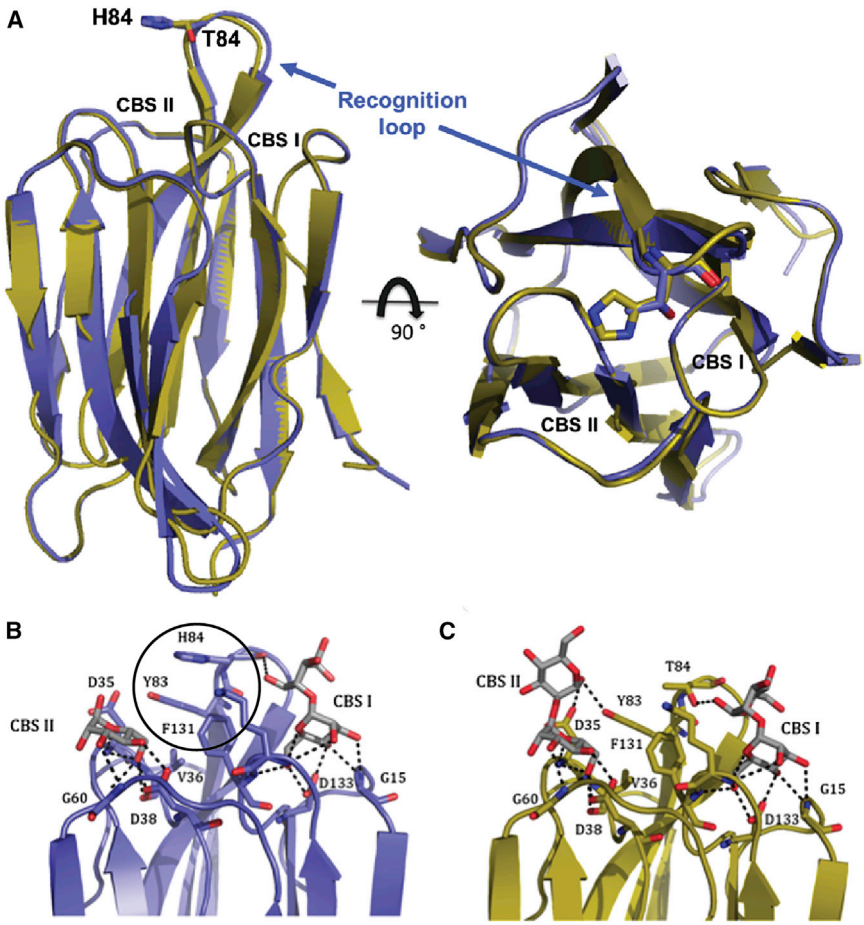


Figure 4. A Comparison of the Crystal Structures of Recombinant WT BanLec and Its H84T Mutant

(A) Overlay of the structures of a monomer of recombinant WT (blue) and H84T (yellow) BanLec. Both structures are presented as cartoons with residue 84 shown in ball and stick with oxygen atoms in red, nitrogen atoms in blue, and carbon atoms in the color of the monomer. The N and C termini are labeled. The right image is the result of rotating the left image 90° toward the viewer. CBS, carbohydrate-binding site.

(B and C) Binding of a dimannoside to WT BanLec in blue (B) and to the H84T mutant in yellow (C). A disaccharide is shown in gray, and individual atoms are colored as in (A). Residues involved in hydrogen bonding are shown in ball and stick, and hydrogen bonds are shown as dashed lines. The pi-pi stacking between Y83 and H84 in the WT protein is circled.

See also Figure S3 and Tables S4 and S5.

visible electron density in CBS I for three out of four chains of the WT protein and all four chains of the H84T protein, but is present in the CBS II for only one H84T chain. For the CBS I site, each protein makes one H-bond with the second mannose moiety. In WT, the H84 side chain does not engage in H-bonds with the second mannose moiety in the CBS I pocket (Figure 4B), while in H84T, the side chain of T84 swings into the CBS I pocket to form a H-bond with this O1 hydroxyl oxygen of the sugar (Figure 4C). The existence of pi-pi stacking locks the imidazole ring of H84 toward the CBS II, and its loss in H84T allows for this reorientation toward the CBS I. Thus, although the global structures of WT and H84T are not markedly different, the loss of pi-pi stacking alters the carbohydrate-protein contacts and topological presentation of the carbohydrate-binding site, potentially explaining the difference in their biological behavior.

NMR Spectroscopy and Molecular Dynamics Simulations Reveal Differences in the Structures of WT and H84T BanLec

We used solution-state NMR spectroscopy to further delineate any differences between WT and H84T. NMR spectra showing a single set of resonances for the monomeric subunit are consistent with both WT and H84T forming symmetric oligomers. However, both proteins exhibited a tendency to aggregate over time and this precluded application of multidimensional NMR experiments for resonance assignments (Sattler et al., 1999). Although BanLec is a dimer in solution at physiological pH (Khan et al., 2013), the X-ray structures reveal the possibility of BanLec forming asymmetric tetramers. Therefore it is probable that the high protein concentration used in NMR promotes the formation of higher-order aggregates. To reduce this tendency, we introduced two mutations: Y46K to disrupt the protein-protein interactions of the tetramer and V66D to increase protein hydrophilicity and to disrupt an additional crystal packing site. The resultant Y46K/V66D mutants of WT and H84T indeed formed stable dimers as judged by ^{15}N NMR spin relaxation measurements (see below) and resulted in spectra very similar to those of BanLec without the Y46K/V66D mutations, with the differences primarily localized around the mutation site (Figures S4A and S5). The double-mutant version of the WT was used to obtain assignments, which were then used to assign its H84T counterpart and the corresponding BanLec proteins lacking the double mutation (Figures S4, S5, and S6). In agreement with the crystal structures, we observed significant overlap when comparing the 2D ^{15}N-^{1}H HSQC spectra of WT and H84T, indicating that the two proteins adopt a similar fold (Figure 5A). However, significant differences in chemical shifts were observed in the third Greek key, indicating that the H84T mutation does affect the structural and/or dynamic properties at this site.

The chemical shift differences between WT and H84T span the entire ligand recognition loop (residues 83–88), which plays important roles in determining the carbohydrate-binding specificity (Figures 5B and S4B). The mutation may broadly affect the conformation of this loop, possibly due to loss of pi-pi stacking as observed in the X-ray structure. We did not observe significant differences in the ^{15}N NMR spin relaxation rates (Palmer,

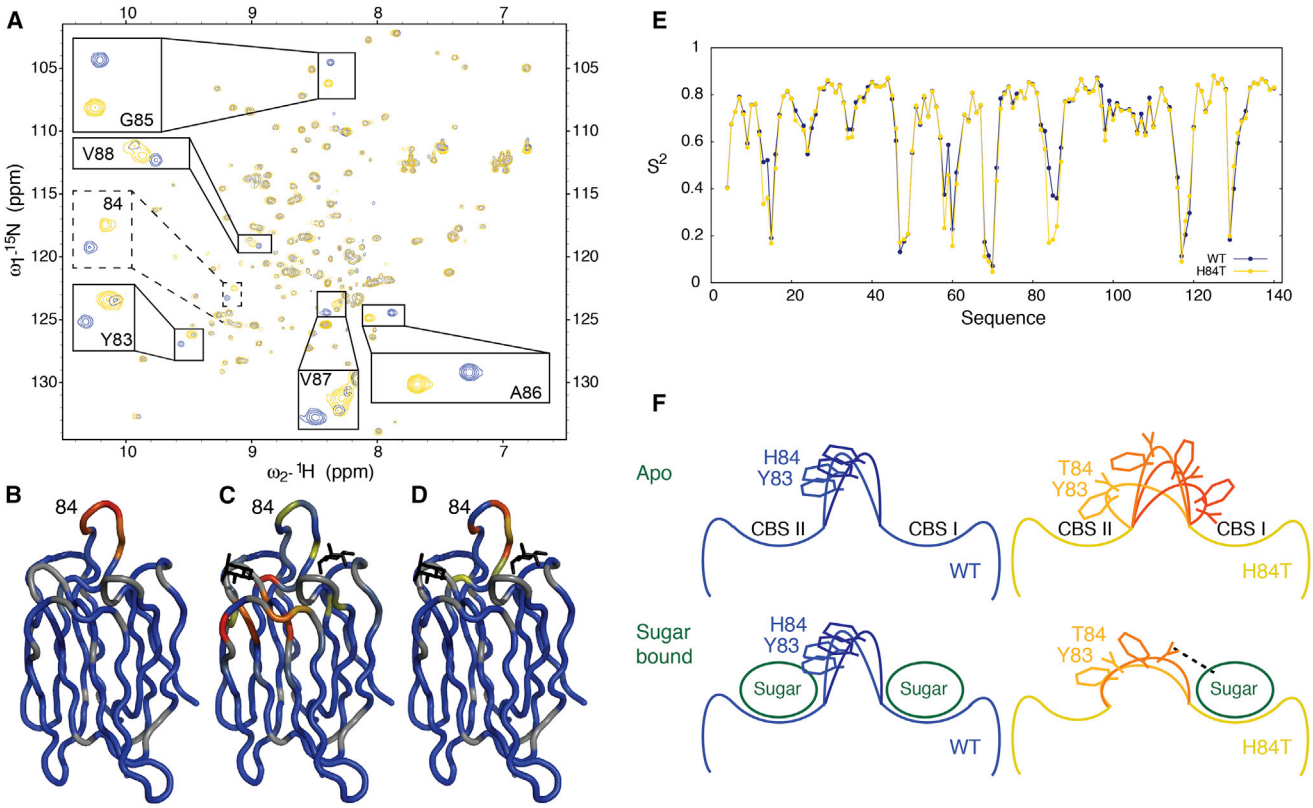


Figure 5. Solution NMR Spectroscopy and Molecular Dynamics Simulations Reveal Dynamic Differences in the Conformations of WT and H84T BanLec at the Third Greek Key

(A) Comparison of H84T mutant and WT BanLec. ^{15}N-^{1}H HSQC spectra of WT (blue) and H84T BanLec (yellow).

(B) Chemical shift changes induced by the H84T mutation color-coded on the structure of WT BanLec.

(C) Chemical shift changes upon pentamannoside binding color-coded on the structure of WT BanLec.

(D) Chemical shift differences between H84T and WT BanLec when interacting with sugar color-coded on the structure of WT BanLec. For (B), (C), and (D), the magnitude in chemical shift change increases from blue (no change) to red (maximal change). Gray corresponds to residues for which the change could not be accurately measured. Sugar moieties are in black.

(E) Comparison of WT and H84T Lipari-Szabo order parameters (S^2 varies between zero and one for maximal to minimal flexibility/amplitude of motions, respectively) computed for WT (blue) and H84T (yellow) using accelerated MD.

(F) Proposed mechanism for separating antiviral activity and mitogenicity using the H84T mutation. Top: in the apo-form (left), the pi-pi stacking interaction helps separate two binding pockets that can engage with branched N-glycans or sugar moieties on different glycan molecules, creating multivalent interactions, while in the H84T mutant loss of pi-pi stacking between residues 84 and 83 results in a more open binding pocket that can engage multiple sugar moieties on the same glycan molecule, limiting the possibility for multivalent interactions. The dashed line symbolizes the capability of the H84T side chain to interact with a sugar in the CBS I, which helps to retain the capability to interact with a single sugar, while mixing the recognition elements of the two binding sites.

See also Figures S4, S5, S6, and S7.

2004) for these and other sites, indicating that WT and H84T have similar dynamics at the picosecond to nanosecond timescales, as well as similar oligomerization states (Figure S7A). This is consistent with the similar dynamics observed for WT and H84T using conventional molecular dynamics (MD) simulations (Figure S7B). However, accelerated MD simulations (Markwick and McCammon, 2011), which can probe slower motions, showed higher flexibility in the ligand recognition loop (83–87) in H84T as compared to the WT protein, consistent with loss of stabilizing pi-pi stacking interactions (Figure 5E).

Next, we performed NMR chemical shift titrations to investigate the interaction of WT and H84T proteins with di- and pentamannosides in solution. The addition of dimannose to WT and H84T or pentamannose to their Y46K/V66D mutant versions resulted in significant chemical shift perturbations or broadening of resonances for residues in and around the sugar-binding pocket defined by the X-ray structure (Figures 5C, S4C, and S4D). In all cases, several resonances from residues involved in sugar binding disappear, e.g., K130 and F131, probably due to exchange broadening (Palmer, 2004). While the sites that experience chemical shift perturbations are very similar between WT and H84T, the perturbations are slightly larger for H84T and differ in direction, particularly for the recognition loop and when binding to pentamannose (Figures 5D, S4C, and S4D). Interestingly, the pentamanose-induced perturbations at 84, 85, and 86 tend to diminish the differences at these sites observed in the absence of sugar, suggesting that sugar binding stabilizes a more similar backbone conformation for these sites (Figures

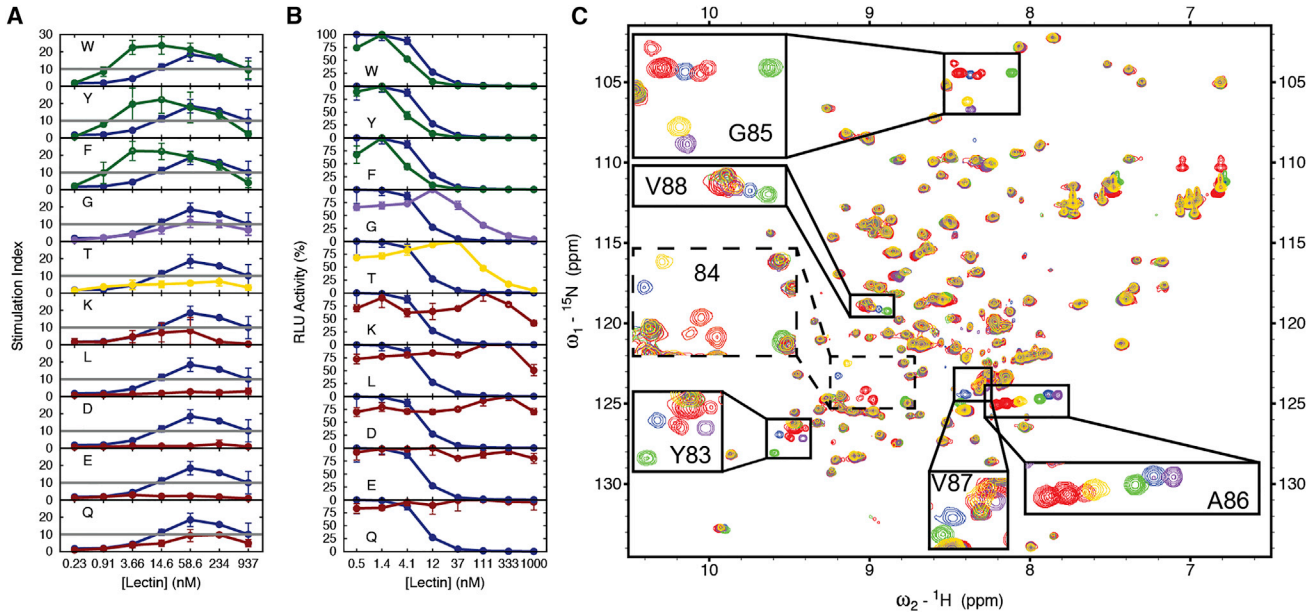


Figure 6. Specific NMR Shifts Correlate with Mitogenicity

(A) Comparison of the mitogenic activity of ten types of H84X mutants to WT BanLec. PBLs were treated with lectin for 3 days and tested for mitogenic activity by the incorporation of BrdU reported from an ELISA in relative luminescent units (RLUs). A stimulation index (RLUs of treated/RLUs of untreated) of less than ten (gray line) is considered non-mitogenic. The type of specific amino-acid substitution at position 84 for each mutant is indicated in each figure. Results with WT are plotted in blue for each comparison shown.

(B) Antiviral activity of the same BanLec mutants. The anti-HIV activity of each BanLec variant was determined by its ability to block infection of TZM-bl cells with virus pseudotyped with the envelope from the HIV-1 BaL strain. The percentage of relative light unit (RLU) activity with increasing concentrations of lectin is plotted for each H84X mutation.

(C) Comparison of NMR chemical shifts for ten representative H84X mutants of BanLec. ^{15}N-^{1}H HSQCs of WT BanLec and the different mutants. The color coding is as follows: WT BanLec (blue), H84T (yellow), H84G (purple), aromatic mutants H84F, H84W, and H84Y (green), and non-aromatic mutants H84D, H84E, H84K, H84Q, and H84L (red).

S4E and S4F). Overall, these data suggest a greater degree of conformational reorganization upon sugar binding in H84T as compared to WT and different sugar-binding modes for the two proteins, consistent with the X-ray structure.

Correlation between Y83-H84 Stacking, NMR Chemical Shifts, Mitogenicity, and Antiviral Activity

To further explore the correlation between Y83-H84 stacking, BanLec conformation, and biological activity, we systematically substituted H84 with amino acids that have different abilities to engage in stacking interactions and then examined the consequence on both NMR spectra and biological activity. A panel of ten H84 BanLec mutants was constructed, systematically replacing the imidazole ring with other aromatic structures or with ionic, polar, or aliphatic groups, and even a hydrogen atom in H84G. These studies employed the version of BanLec without the double Y46K/V66D mutation. Replacing H84 with the aromatic residues tryptophan (H84W), tyrosine (H84Y), and phenylalanine (H84F), which can maintain favorable stacking interactions, had minimal effects on mitogenicity and anti-HIV activity (Figures 6A and 6B). In contrast, replacement of the imidazole by ionic, polar, or aliphatic side chains, including substitutions by the amino acids lysine (H84K), aspartic acid (H84D), glutamic acid (H84E), glutamine (H84Q), and leucine (H84L), resulted in the marked loss of both mitogenicity and anti-HIV activity (Figures 6A and 6B). Only a single mutation (H84G) in this panel of protein variants yielded a reasonably similar, but smaller, drop in mitogenicity as did H84T, while preserving some antiviral HIV activity.

NMR spectra of the different mutants yielded excellent overall overlap, indicating that they all adopt a similar protein fold. The differences relative to WT protein were concentrated in the third Greek key (83–87) (Figure 6C). Further analysis of these differences yielded an interesting trend for several residues; for A86, the resonances observed in all the mutants fall roughly along a straight line. Similar behaviors, too, were observed for V87 and V88, though the magnitude of the change is smaller and more difficult to resolve due to spectral overlap. Furthermore, mutants with aromatic residues (H84F, H84W, and H84Y) that can support pi-pi stacking between residues 83–84 and that have higher mitogenicity and anti-HIV activity have A86 resonance clustering upfield along the line as compared to other mutants that disrupt pi-pi stacking and have lower mitogenicity and reduced anti-HIV activity (Figure 6). Interestingly, in H84G, which exhibits mitogenicity, A86 also clusters upfield with the other mitogenic mutants despite disruption of pi-pi stacking. A simple explanation is that BanLec exists in rapid dynamic equilibrium between two states and that the mutations differentially shift the relative population of the two states, with A86, situated on the opposite side of the third Greek key loop relative to 84, acting as a reporter for

this equilibrium shift. For V87 and V88, such trends are more difficult to discern, but clearly resonances cluster depending on whether aromatic or non-aromatic residues are used in the substitution. These two residues constitute the back part of the third Greek key loop. These results suggest that the mutants with aromatic residues maintain pi-pi stacking interactions between amino acids 83 and 84 (Figure 5F).

Unlike other mutants, H84T retains antiviral activity despite the loss of mitogenicity. Interestingly, in H84T, the A86 resonance presents intermediate NMR characteristics between aromatic and non-aromatic mutants, whereas V87 and V88 cluster with non-aromatic residues. Additionally, in H84T G85 presents a very distinct signature, shared only with the H84G, which also retains some antiviral activity, indicating that the two mutants share some unique conformational properties. This suggests that the third Greek key loop in the H84T mutant uniquely combines conformational attributes of aromatic and non-aromatic mutants.

Molecular Basis for Separating Two Activities of BanLec

Both mitogenicity and antiviral activities of BanLec require association with N-glycans, and so most mutations that block mitogenicity also abolish antiviral activity (Figures 1A and 6). Unlike other mutants, H84T retains high antiviral activity, which requires the capacity to home in on viral glycoproteins with sufficient affinity. This is achieved despite disrupting pi-pi Y83-H84 stacking, which is important for sugar binding, possibly due to compensatory interactions between the side chain of T84 and the sugar and retention of WT-like conformational properties.

In contrast, mitogenicity requires the ability to cross-link cognate binding partners, beyond a simple association. Our data suggest that the loss of 83-84 stacking decreases this capacity in H84T and other mutants both due to slightly reduced sugar binding affinity (and possibly altered sugar binding specificity) and also due to disruption of the wall that helps create two independent sugar-binding sites, each capable of interacting with a distinct glycan molecule (Figure 5F, left). Rather, in H84T, T84 rotates away from CBS II to interact with sugars in CBS I, effectively mixing recognition elements in the two binding sites (Figure 5F, right). This more open binding pocket may make it more likely for the same glycan molecule to simultaneously interact with the two sugar-binding sites and/or binding at one site may engage elements from the second site, resulting in weaker binding affinity for a second glycan molecule (Figure 5F, right). This makes it less likely for H84T to simultaneously interact with multiple glycan molecules as required for mitogenicity.

DISCUSSION

The Sugar Code underlies a key biological route of information transfer by which cell-to-cell interactions and cell signaling are orchestrated. Indeed, sugars can be considered the third type of biological alphabet, along with nucleotides and amino acids (Murphy et al., 2013). The receptors for glycans (lectins) are endowed with the capacity to target distinct counterreceptors by their structure and topological mode of presentation (Gabius et al., 2015). In doing so, lectins can play a vital role in regulating biological processes, such as cell growth and the immune response, and also serve as tools for studying structural aspects of glycobiology (Kaltner and Gabius, 2012).

It has previously been observed that a single sugar unit can act as a switch for a complex-type glycan's 3D structure, thus altering its ligand reactivity and subsequent signaling (Gabius et al., 2011). In the case of a bacterial lectin, the H57A substitution in the cholera toxin B-subunit did not disrupt binding to the GM1 ganglioside, but did lead to loss of immunomodulatory activity and the ability to induce apoptosis, with altered loop position and rigidification affecting further cell surface contacts (Aman et al., 2001). SNPs occur naturally in the genes of human and animal lectins, and these natural sequence changes can affect the carbohydrate recognition domain and biological function, as seen with a human galactose-binding lectin (Ruiz et al., 2014). In this latter case, an impact on cell proliferation and *trans*-interactions has been inferred (Ruiz et al., 2014; Zhang et al., 2015a). Here, we have demonstrated that two distinct properties of a lectin can be separated through rational molecular fine-tuning: BanLec can be engineered to essentially lose its mitogenicity while retaining very potent antiviral activity. The resultant H84T BanLec mutant is a broad-spectrum antiviral agent that is highly active against multiple strains of HCV, influenza, and HIV-1 in tissue culture and in vivo; it will also likely prove effective against other clinically important viruses with a suitable presentation of mannose on their surfaces.

Our data suggest that loss of mitogenicity can be achieved by disrupting 83-84 stacking and disrupting a wall separating two sugar-binding pockets, thus diminishing polyvalent interactions. However, doing so while retaining antiviral activity requires a specific amino-acid substitution (H84T) that may help retain WT conformational properties, as well as possibly form unique contacts that can compensate for loss of interactions with the 83-84 stack. It is possible that these basic design principles can be applied and extended to allow rational engineering of other lectins for use as antiviral tools and other therapeutic purposes. The recent demonstration that *trans*-interactions can be strengthened by the insertion of a linker into the homodimer of the antiviral galectin-1 (Zhang et al., 2015b) and the work presented here encourage such efforts. While the term lectin etymologically stems from the Latin word "legere," meaning to pick, choose, or select (Boyd, 1954), thus emphasizing the natural ability of these proteins to target specific carbohydrates, we have shown that lectins can be made yet more selective through molecular engineering. Our findings also suggest that custom-designed lectins can be employed to tease apart fine mechanisms of immune activation. In more general terms, this proof-of-principle work is likely to inspire the generation of new and innovative tools in the quest to delineate the intricacies of the Sugar Code.

EXPERIMENTAL PROCEDURES

Construction and Mutation of BanLec Expression Vectors and Purification of Recombinant BanLec Mutants

The BanLec cDNA was cloned into a vector, allowing for expression of His-tagged protein in *E. coli*, mutagenesis, and purification over a nickel column as described in the Supplemental Experimental Procedures.

Assessment of Anti-HIV Activity

Assays testing the anti-HIV activity of WT and H84T BanLec in PBMCs were performed as described previously, measuring p24 for HIV-1 and p27 for HIV-2 (Férir et al., 2011). For the TZM-bl cell assays, to each well of a white 96-well plate 100 μl of a solution containing cells, resuspended at 1×10^5 cells/ml in DMEM medium with 25 mM HEPES and 10% FBS, was added. The next day, the medium was removed by aspiration and fresh medium containing lectin or PBS as a control was added to the plate at a concentration 2-fold higher than the final concentration. After 30 min of incubation, virus diluted with medium was added, and the cells were incubated for 48 hr at 37°C. After the incubation, 100 μl of medium were removed and replaced with 100 μl of ONE-Glo Luciferase reagent (Promega) for determination of luciferase expression.

HCV Experiments

The anti-HCV activity of BanLec derivatives was determined for different genotypic chimeras in Huh-7.5 cells using bicistronic *Gaussia* luciferase reporter genomes as described in the Supplemental Experimental Procedures.

Assessment of Anti-Influenza Activity

The in vitro anti-influenza activity of H84T and its efficacy when administered via the intranasal route to female BALB/c mice challenged with influenza were assessed as described in the Supplemental Experimental Procedures.

Hemagglutination Assay and ITC

Hemagglutination assays conducted using rabbit erythrocytes and ITC were carried out as described in the Supplemental Experimental Procedures.

Assessment of Mitogenic Activity by BrdU Incorporation

Mitogenic activity was quantified as is described in the legend of Figure 6 and further in the Supplemental Experimental Procedures.

Flow Cytometry to Measure Cellular Activation and Bio-Plex Cytokine Assay

Expression of CD69 was measured by flow cytometry and cytokine production following stimulation with lectin by Bio-Plex assay as described in the Supplemental Experimental Procedures.

Vaginal HIV-1 Transmission

BLT mice were anesthetized and received 75 μg of H84T BanLec vaginally in a volume of 20 μl. 10 min after application of the lectin, the mice were challenged vaginally with 175,000 TCIU of HIV-1 JR-CSF. Mice were bled weekly and the plasma was analyzed for the presence of viral RNA for 6 weeks as described previously (Wahl et al., 2012).

Glycocluster Synthesis and Assays

Synthesis of the glycoclusters is described in the Supplemental Experimental Procedures. The determination of the relative ability of glycoclusters to inhibit lectin binding to a matrix presenting a glycoligand, given as the inhibitory concentration (IC) at which the spectrophotometrically determined signal intensity is reduced by 50% (IC_{50} value), provides a measure of the engagement of a lectin in multivalent associations. This value and the sensitivity of lectin binding to the surface of cells in culture in the presence of glycoclusters were assayed as described in the Supplemental Experimental Procedures.

NMR Spectroscopy

All NMR experiments were acquired at 313 K on a 600 MHz spectrometer equipped with a triple-resonance cryoprobe. Y46K/V66D BanLec assignment was obtained using a classical 3D assignment strategy. For a more detailed description, see Supplemental Experimental Procedures.

Crystallization, Data Collection, and Structure Determination

Following crystallization, data were obtained by LS-CAT, and structure, in the presence or absence of dimannoside, was determined as noted in the Supplemental Experimental Procedures.

MD Simulations

MD simulations were conducted as described in the Supplemental Experimental Procedures. All simulations were conducted using the Amber 12 package (Case et al., 2005) with the ff99SB*-ILDN force field (Hornak et al., 2006; Lindorff-Larsen et al., 2012). The accelerated MD simulations were set up following published protocols (Pierce et al., 2012).

ACCESSION NUMBERS

The accession number for the crystal structure reported in this paper is deposited in PDB: 3RFP. The accession numbers for wild-type BanLec, wild-type in complex with dimannoside, H84T BanLec mutant, and H84T BanLec mutant in complex with dimannoside, respectively, reported in this paper are deposited in PDB: 4PIF, 4PIK, 4PIT, 4PIU.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.056.

AUTHOR CONTRIBUTIONS

M.D.S. created the H84T mutant, tested its activity against HIV in vivo and in vitro, assessed mitogenicity, and helped write the manuscript. D.M.B. created the multiple H84 mutants, produced the BanLec to be analyzed by NMR, and ran NMR assays. L.S. performed the NMR studies, along with J.C., determined the NMR assignments, and helped write the manuscript. H.C.W. and I.J.G. performed agglutination assays and ITC. M.H.K. suggested separating mitogenicity from antiviral activity and helped design experiments. P.V.M., S.O., and R.R. designed and synthesized the glycoclusters. S.K. performed antiviral studies and refined the production of BanLec. E.B.T., B.L.H., and D.F.S. determined the activity of H84T against WT influenza in vitro and in vivo. C.F., H.-H.H., and C.M.R. devised and conducted the HCV experiments. D.S. further determined the activity of H84T against HIV and tested parameters of mitogenicity. J.V.G. helped design and supervised the in vivo HIV studies. J.L.M. and J.A.S. performed the X-ray crystallography studies. S.A. and H.-J.G. designed and carried out the studies using glycoclusters and played a major role in writing the manuscript. Y.X. performed the MD simulations. H.M.A-H. directed all structural studies and, along with D.M.M., supervised and integrated the work done by the collaborators and wrote the manuscript.

ACKNOWLEDGMENTS

The authors are grateful to Evelyn Coves-Datson, Anjan Saha, Dana Huskens, Jen Lewis, and Dr. Derek Dube for assistance, Dr. David Smith of LS-CAT for help with remote data collection, and Drs. B. Friday and A. Leddoz for inspiring discussions. Work in the laboratories of D.M.M. and H.M.A-H. was supported by an NIH grant (1R01CA144043). H.-J.G. was supported by the EC-funded GlycoHIT program (contract no. 260600) and Training Network GLYCOPHARM (PITN-GA-2012-317297). M.D.S. and J.V.G. were supported by grants from the NIH (AI096138, AI073146, and P30 AI50410). P.V.M. has been supported by Marie Curie Intra-European Fellowships (500748, 514958, and 220948), the Programme for Research in Third-Level Institutions (PRTLI), administered by the Higher Education Authority, the Irish Research Council, Enterprise Ireland, and Science Foundation Ireland (04/BR/C0192, 06/RFP/CHO032, and 12/IA/1398). R.R. is grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) for financial support and for a Canadian Research Chair in Therapeutic Chemistry. The participation of A. Papadopoulos and T.C. Shiao is also acknowledged in the preparation of compounds 4–8. M.H.K. received support from the Concerned Parents for AIDS Research. D.S. was supported by KU Leuven grants (GOA 10/014 and PF 10/18), a European CHAARM grant (242135), and an equipment grant from the Fondation Dormeur, Vaduz. Work in the laboratory of C.M.R. was supported in part by PHS grants (R01 AI099284, R01 AI072613, and R01

CA057973). Work at the Utah State University was supported by a grant (contract number HHSN27220100000391/HHSN27200005/A37) from the Respiratory Diseases Branch, Division of Microbiology and Infectious Diseases, NIAID, NIH. J.A.S., J.L.M, and H.M.A.-H. were partially supported by a grant (P50 GM103297) from the NIH. J.A.S. and J.L.M were also supported in part by the University of Michigan Center for Structural Biology. Use of the Advanced Photon Source was funded by the U.S. Department of Energy (under contract no. DE-AC02-06CH11357), and use of the LS-CAT Sector 21 was funded by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (085P1000817). D.M.M. is the founder of Virule, a company formed to commercialize H84T BanLec. This work is dedicated to the memory of Dr. John Hilfinger, who taught D.M.M. to appreciate biochemistry.

Received: July 16, 2014
Revised: June 5, 2015
Accepted: September 29, 2015
Published: October 22, 2015

REFERENCES

Aman, A.T., Fraser, S., Merritt, E.A., Rodigherio, C., Kenny, M., Ahn, M., Hol, W.G., Williams, N.A., Lencer, W.I., and Hirst, T.R. (2001). A mutant cholera toxin B subunit that binds GM1 ganglioside but lacks immunomodulatory or toxic activity. Proc. Natl. Acad. Sci. USA *98*, 8536–8541.

André, S., Kaltner, H., Manning, J.C., Murphy, P.V., and Gabius, H.-J. (2015). Lectins: getting familiar with translators of the sugar code. Molecules *20*, 1788–1823.

Borrebaeck, C.A.K., and Carlsson, R. (1989). Lectins as mitogens. Adv. Lectin Res. *2*, 1–27.

Boyd, W.C. (1954). The proteins of immune reactions. In The Proteins, H. Neurath and K. Bailey, eds. (New York: Academic Press), pp. 756–844.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J. Comput. Chem. *26*, 1668–1688.

Férir, G., Vermeire, K., Huskens, D., Balzarini, J., Van Damme, E.J.M., Kehr, J.C., Dittmann, E., Swanson, M.D., Markovitz, D.M., and Schols, D. (2011). Synergistic in vitro anti-HIV type 1 activity of tenofovir with carbohydrate-binding agents (CBAs). Antiviral Res. *90*, 200–204.

Gabius, H.-J. (2015). The magic of the sugar code. Trends Biochem. Sci. *40*, 341.

Gabius, H.-J., André, S., Jiménez-Barbero, J., Romero, A., and Solís, D. (2011). From lectin structure to functional glycomics: principles of the sugar code. Trends Biochem. Sci. *36*, 298–313.

Gabius, H.-J., Kaltner, H., Kopitz, J., and André, S. (2015). The glycobiology of the CD system: a dictionary for translating marker designations into glycan/lectin structure and function. Trends Biochem. Sci. *40*, 360–376.

Gavrovic-Jankulovic, M., Poulsen, K., Brckalo, T., Bobic, S., Lindner, B., and Petersen, A. (2008). A novel recombinantly produced banana lectin isoform is a valuable tool for glycoproteomics and a potent modulator of the proliferation response in CD3+, CD4+, and CD8+ populations of human PBMCs. Int. J. Biochem. Cell Biol. *40*, 929–941.

Goffard, A., Callens, N., Bartosch, B., Wychowski, C., Cosset, F.L., Montpellier, C., and Dubuisson, J. (2005). Role of N-linked glycans in the functions of hepatitis C virus envelope glycoproteins. J. Virol. *79*, 8400–8409.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins *65*, 712–725.

Huskens, D., Vermeire, K., Vandemeulebroucke, E., Balzarini, J., and Schols, D. (2008). Safety concerns for the potential use of cyanovirin-N as a microbicidal anti-HIV agent. Int. J. Biochem. Cell Biol. *40*, 2802–2814.

Kaltner, H., and Gabius, H.-J. (2012). A toolbox of lectins for translating the sugar code: the galectin network in phylogenesis and tumors. Histol. Histopathol. *27*, 397–416.

Khan, J.M., Qadeer, A., Ahmad, E., Ashraf, R., Bhushan, B., Chaturvedi, S.K., Rabbani, G., and Khan, R.H. (2013). Monomeric banana lectin at acidic pH overrules conformational stability of its native dimeric form. PLoS ONE *8*, e62428.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E. (2012). Systematic validation of protein force fields against experimental data. PLoS ONE *7*, e32131.

Markwick, P.R., and McCammon, J.A. (2011). Studying functional dynamics in bio-molecules using accelerated molecular dynamics. Phys. Chem. Chem. Phys. *13*, 20053–20065.

Meagher, J.L., Winter, H.C., Ezell, P., Goldstein, I.J., and Stuckey, J.A. (2005). Crystal structure of banana lectin reveals a novel second sugar binding site. Glycobiology *15*, 1033–1042.

Mo, H., Winter, H.C., van Damme, E.J., Peumans, W.J., Misaki, A., and Goldstein, I.J. (2001). Carbohydrate binding properties of banana (*Musa acuminata*) lectin I. Novel recognition of internal α1,3-linked glucosyl residues. Eur. J. Biochem. *268*, 2609–2615.

Murphy, P.V., André, S., and Gabius, H.-J. (2013). The third dimension of reading the sugar code by lectins: design of glycoclusters with cyclic scaffolds as tools with the aim to define correlations between spatial presentation and activity. Molecules *18*, 4026–4053.

Nakamura-Tsuruta, S., Uchiyama, N., Peumans, W.J., Van Damme, E.J.M., Totani, K., Ito, Y., and Hirabayashi, J. (2008). Analysis of the sugar-binding specificity of mannose-binding-type Jacalin-related lectins by frontal affinity chromatography–an approach to functional classification. FEBS J. *275*, 1227–1239.

Ng, W.C., Tate, M.D., Brooks, A.G., and Reading, P.C. (2012). Soluble host defense lectins in innate immunity to influenza virus. J. Biomed. Biotechnol. *2012*, 732191.

Palmer, A.G., 3rd. (2004). NMR characterization of the dynamics of biomacromolecules. Chem. Rev. *104*, 3623–3640.

Pierce, L.C., Salomon-Ferrer, R., Augusto, F.d.O.C., McCammon, J.A., and Walker, R.C. (2012). Routine access to millisecond time scale events with accelerated molecular dynamics. Chem. Theory Comput. *8*, 2997–3002.

Reyes-del Valle, J., de la Fuente, C., Turner, M.A., Springfeld, C., Apte-Sengupta, S., Frenzke, M.E., Forest, A., Whidby, J., Marcotrigiano, J., Rice, C.M., and Cattaneo, R. (2012). Broadly neutralizing immune responses against hepatitis C virus induced by vectored measles viruses and a recombinant envelope protein booster. J. Virol. *86*, 11558–11566.

Ruiz, F.M., Scholz, B.A., Buzamet, E., Kopitz, J., André, S., Menéndez, M., Romero, A., Solís, D., and Gabius, H.-J. (2014). Natural single amino acid polymorphism (F19Y) in human galectin-8: detection of structural alterations and increased growth-regulatory activity on tumor cells. FEBS J. *281*, 1446–1464.

Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog. NMR Spectr. *34*, 93–158.

Singh, D.D., Saikrishnan, K., Kumar, P., Surolia, A., Sekar, K., and Vijayan, M. (2005). Unusual sugar specificity of banana lectin from *Musa paradisiaca* and its probable evolutionary origin. Crystallographic and modelling studies. Glycobiology *15*, 1025–1032.

Singh, S.S., Devi, S.K., and Ng, T.B. (2014). Banana lectin: a brief review. Molecules *19*, 18817–18827.

Smee, D.F., Bailey, K.W., Wong, M.H., O'Keefe, B.R., Gustafson, K.R., Mishin, V.P., and Gubareva, L.V. (2008). Treatment of influenza A (H1N1) virus infections in mice and ferrets with cyanovirin-N. Antiviral Res. *80*, 266–271.

Solís, D., Bovin, N.V., Davis, A.P., Jiménez-Barbero, J., Romero, A., Roy, R., Smetana, K., Jr., and Gabius, H.-J. (2015). A guide into glycosciences: how chemistry, biochemistry and biology cooperate to crack the sugar code. Biochim. Biophys. Acta *1850*, 186–235.

Swanson, M.D., Winter, H.C., Goldstein, I.J., and Markovitz, D.M. (2010). A lectin isolated from bananas is a potent inhibitor of HIV replication. J. Biol. Chem. *285*, 8646–8655.

Wahl, A., Swanson, M.D., Nochi, T., Olesen, R., Denton, P.W., Chateau, M., and Garcia, J.V. (2012). Human breast milk and antiretrovirals dramatically reduce oral HIV-1 transmission in BLT humanized mice. PLoS Pathog. *8*, e1002732.

Winter, H.C., Oscarson, S., Slättegård, R., Tian, M., and Goldstein, I.J. (2005). Banana lectin is unique in its recognition of the reducing unit of 3-O-β-glucosyl/mannosyl disaccharides: a calorimetric study. Glycobiology *15*, 1043–1050.

Zhang, S., Moussodia, R.-O., Vértesy, S., André, S., Klein, M.L., Gabius, H.-J., and Percec, V. (2015a). Unraveling functional significance of natural variations of a human galectin by glycodendrimersomes with programmable glycan surface. Proc. Natl. Acad. Sci. USA *112*, 5585–5590.

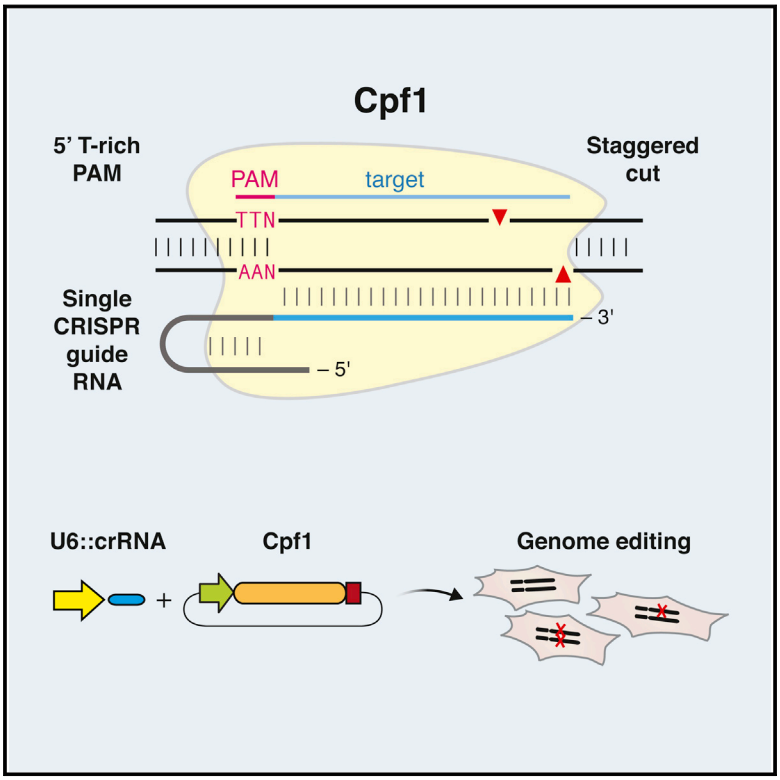
Zhang, S., Moussodia, R.-O., Murzeau, C., Sun, H.J., Klein, M.L., Vértesy, S., André, S., Roy, R., Gabius, H.-J., and Percec, V. (2015b). Dissecting molecular aspects of cell interactions using glycodendrimersomes with programmable glycan presentation and engineered human lectins. Angew. Chem. Int. Ed. Engl. *54*, 4036–4040.

Article

Cell

Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System

Graphical Abstract



Authors

Bernd Zetsche, Jonathan S. Gootenberg, Omar O. Abudayyeh, ..., Aviv Regev, Eugene V. Koonin, Feng Zhang

Correspondence

zhang@broadinstitute.org

In Brief

Cpf1 is a RNA-guided DNA nuclease that provides immunity in bacteria and can be adapted for genome editing in mammalian cells.

Highlights

- CRISPR-Cpf1 is a class 2 CRISPR system
- Cpf1 is a CRISPR-associated two-component RNA-programmable DNA nuclease
- Targeted DNA is cleaved as a 5-nt staggered cut distal to a 5′ T-rich PAM
- Two Cpf1 orthologs exhibit robust nuclease activity in human cells

Zetsche et al., 2015, Cell *163*, 759–771
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.038

CellPress

Cell

Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System

Bernd Zetsche,[1,2,3,4,5,10] Jonathan S. Gootenberg,[1,2,3,4,6,10] Omar O. Abudayyeh,[1,2,3,4] Ian M. Slaymaker,[1,2,3,4] Kira S. Makarova,[7] Patrick Essletzbichler,[1,2,3,4] Sara E. Volz,[1,2,3,4] Julia Joung,[1,2,3,4] John van der Oost,[8] Aviv Regev,[1,9] Eugene V. Koonin,[7] and Feng Zhang[1,2,3,4,*]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]McGovern Institute for Brain Research
[3]Department of Brain and Cognitive Sciences
[4]Department of Biological Engineering
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Department of Developmental Pathology, Institute of Pathology, Bonn Medical School, Sigmund Freud Street 25, 53127 Bonn, Germany
[6]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[7]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[8]Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, Netherlands
[9]Department of Biology, Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[10]Co-first author
*Correspondence: zhang@broadinstitute.org
http://dx.doi.org/10.1016/j.cell.2015.09.038

SUMMARY

The microbial adaptive immune system CRISPR mediates defense against foreign genetic elements through two classes of RNA-guided nuclease effectors. Class 1 effectors utilize multi-protein complexes, whereas class 2 effectors rely on single-component effector proteins such as the well-characterized Cas9. Here, we report characterization of Cpf1, a putative class 2 CRISPR effector. We demonstrate that Cpf1 mediates robust DNA interference with features distinct from Cas9. Cpf1 is a single RNA-guided endonuclease lacking tracrRNA, and it utilizes a T-rich protospacer-adjacent motif. Moreover, Cpf1 cleaves DNA via a staggered DNA double-stranded break. Out of 16 Cpf1-family proteins, we identified two candidate enzymes from *Acidaminococcus* and *Lachnospiraceae*, with efficient genome-editing activity in human cells. Identifying this mechanism of interference broadens our understanding of CRISPR-Cas systems and advances their genome editing applications.

INTRODUCTION

Almost all archaea and many bacteria achieve adaptive immunity through a diverse set of CRISPR-Cas (*c*lustered *r*egularly *i*nterspaced *s*hort *p*alindromic *r*epeats and *C*RISPR-*a*ssociated proteins) systems, each of which consists of a combination of Cas effector proteins and CRISPR RNAs (crRNAs) (Makarova et al., 2011, 2015). The defense activity of the CRISPR-Cas systems includes three stages: (1) adaptation, when a complex of Cas proteins excises a segment of the target DNA (known as a protospacer) and inserts it into the CRISPR array (where this sequence becomes a spacer); (2) expression and processing of the precursor CRISPR (pre-cr) RNA resulting in the formation of mature crRNAs; and (3) interference, when the effector module—either another Cas protein complex or a single large protein—is guided by a crRNA to recognize and cleave target DNA (or in some cases, RNA) (Horvath and Barrangou, 2010; Sorek et al., 2013; Barrangou and Marraffini, 2014). The adaptation stage is mediated by the complex of the Cas1 and Cas2 proteins, which are shared by all known CRISPR-Cas systems, and sometimes involves additional Cas proteins. Diversity is observed at the level of processing of the pre-crRNA to mature crRNA guides, proceeding via either a Cas6-related ribonuclease or a housekeeping RNaseIII that specifically cleaves double-stranded RNA hybrids of pre-crRNA and tracrRNA. Moreover, the effector modules differ substantially among the CRISPR-Cas systems (Makarova et al., 2011, 2015; Charpentier et al., 2015). In the latest classification, the diverse CRISPR-Cas systems are divided into two classes according to the configuration of their effector modules: class 1 CRISPR systems utilize several Cas proteins and the crRNA to form an effector complex, whereas class 2 CRISPR systems employ a large single-component Cas protein in conjunction with crRNAs to mediate interference (Makarova et al., 2015).

Multiple class 1 CRISPR-Cas systems, which include the type I and type III systems, have been identified and functionally characterized in detail, revealing the complex architecture and dynamics of the effector complexes (Brouns et al., 2008; Marraffini and Sontheimer, 2008; Hale et al., 2009; Sinkunas et al., 2013; Jackson et al., 2014; Mulepati et al., 2014). Several class 2 CRISPR-Cas systems have also been identified and experimentally characterized, but they are all type II and employ homologous RNA-guided endonucleases of the Cas9 family as effectors (Barrangou et al., 2007; Garneau et al., 2010; Deltcheva et al., 2011; Sapranauskas et al., 2011; Jinek et al., 2012; Gasiunas et al., 2012). A second, putative class 2 CRISPR system, tentatively assigned to type V, has

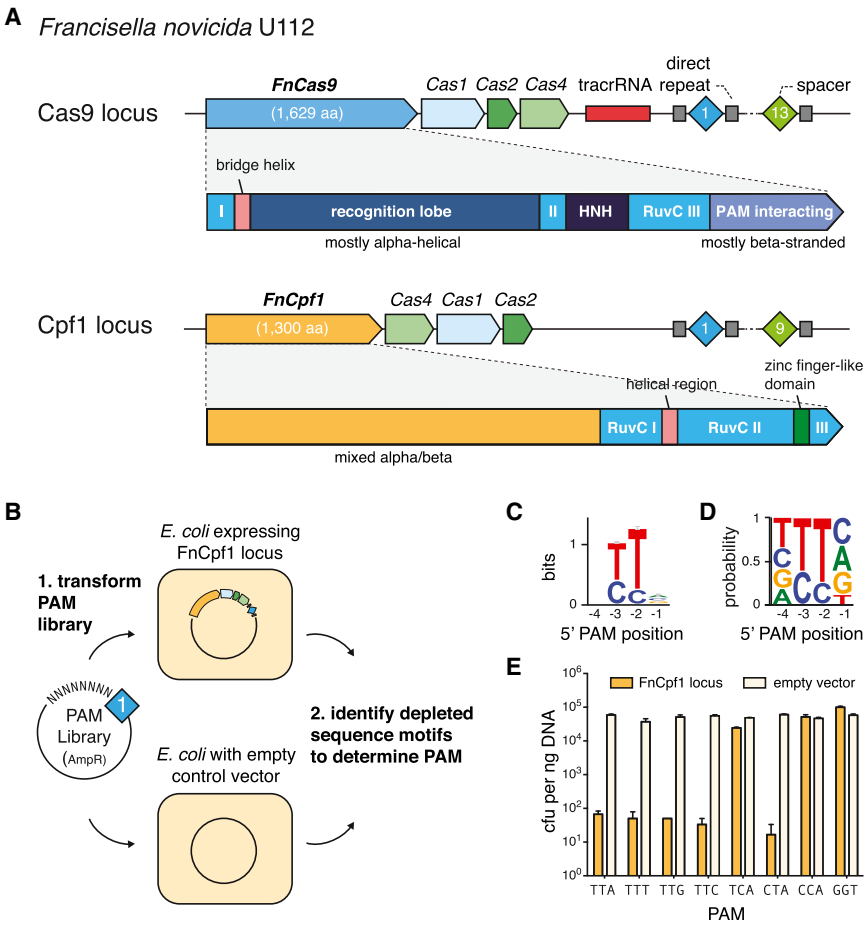


Figure 1. The *Francisella novicida U112* Cpf1 CRISPR Locus Provides Immunity against Transformation of Plasmids Containing Protospacers Flanked by a 5′-TTN PAM

(A) Organization of two CRISPR loci found in *Francisella novicida U112* (NC_008601). The domain architectures of FnCas9 and FnCpf1 are compared.

(B) Schematic illustrating the plasmid depletion assay for discovering the PAM position and identity. Competent *E. coli* harboring either the heterologous FnCpf1 locus plasmid (pFnCpf1) or the empty vector control were transformed with a library of plasmids containing the matching protospacer flanked by randomized 5′ or 3′ PAM sequences and selected with antibiotic to deplete plasmids carrying successfully targeted PAM. Plasmids from surviving colonies were extracted and sequenced to determine depleted PAM sequences.

(C and D) Sequence logo for the FnCpf1 PAM as determined by the plasmid depletion assay. Letter height at each position is measured by information content (C) or frequency (D); error bars show 95% Bayesian confidence interval.

(E) *E. coli* harboring pFnCpf1 provides robust interference against plasmids carrying 5′-TTN PAMs (n = 3; error bars represent mean ± SEM). See also Figure S1.

been recently identified in several bacterial genomes (http://www.jcvi.org/cgi-bin/tigrfams/HmmReportPage.cgi?acc=TIGR04330) (Schunder et al., 2013; Vestergaard et al., 2014; Makarova et al., 2015). The putative type V CRISPR-Cas systems contain a large, ~1,300 amino acid protein called Cpf1 (CRISPR from *Prevotella* and *Francisella* 1). It remains unknown, however, whether Cpf1-containing CRISPR loci indeed represent functional CRISPR systems. Given the broad applications of Cas9 as a genome-engineering tool (Hsu et al., 2014; Jiang and Marraffini, 2015), we sought to explore the function of Cpf1-based putative CRISPR systems.

Here, we show that Cpf1-containing CRISPR-Cas loci of *Francisella novicida U112* encode functional defense systems capable of mediating plasmid interference in bacterial cells guided by the CRISPR spacers. Unlike Cas9 systems, Cpf1-containing CRISPR systems have three features. First, Cpf1-associated CRISPR arrays are processed into mature crRNAs without the requirement of an additional *trans*-activating crRNA (tracrRNA) (Deltcheva et al., 2011; Chylinski et al., 2013). Second, Cpf1-crRNA complexes efficiently cleave target DNA proceeded by a short T-rich protospacer-adjacent motif (PAM), in contrast to the G-rich PAM following the target DNA for Cas9 systems. Third, Cpf1 introduces a staggered DNA double-stranded break with a 4 or 5-nt 5′ overhang.

To explore the suitability of Cpf1 for genome-editing applications, we characterized the RNA-guided DNA-targeting requirements for 16 Cpf1-family proteins from diverse bacteria, and we identified two Cpf1 enzymes from *Acidaminococcus sp. BV3L6* and *Lachnospiraceae bacterium ND2006* that are capable of mediating robust genome editing in human cells. Collectively, these results establish Cpf1 as a class 2 CRISPR-Cas system that includes an effective single RNA-guided endonuclease with distinct properties that has the potential to substantially advance our ability to manipulate eukaryotic genomes.

RESULTS

Cpf1-Containing CRISPR Loci Are Active Bacterial Immune Systems

Cpf1 was first annotated as a CRISPR-associated gene in TIGRFAM (http://www.jcvi.org/cgi-bin/tigrfams/HmmReportPage.cgi?acc=TIGR04330) and has been hypothesized to be the effector of a CRISPR locus that is distinct from the Cas9-containing type II CRISPR-Cas loci that are also present in the genomes of some of the same bacteria, such as multiple strains of *Francisella* and *Prevotella* (Schunder et al., 2013; Vestergaard et al., 2014; Makarova et al., 2015) (Figure 1A). The Cpf1 protein contains a predicted RuvC-like endonuclease domain that is distantly related to the respective nuclease domain of Cas9. However, Cpf1 differs from Cas9 in that it lacks a second, HNH endonuclease domain, which is inserted within the

RuvC-like domain of Cas9. Furthermore, the N-terminal portion of Cpf1 is predicted to adopt a mixed α/β structure and appears to be unrelated to the N-terminal, α-helical recognition lobe of Cas9 (Figure 1A). It has been shown that the nuclease moieties of Cas9 and Cpf1 are homologous to distinct groups of transposon-encoded TnpB proteins, the first one containing both RuvC and HNH nuclease domains and the second one containing the RuvC-like domain only (Makarova and Koonin, 2015). Apart from these distinctions between the effector proteins, the Cpf1-carrying loci encode Cas1, Cas2, and Cas4 proteins that are more closely related to orthologs from types I and III than to those from type II CRISPR systems (Makarova et al., 2015). Taken together, these differences from type II have prompted the classification of Cpf1-encoding CRISPR-Cas loci as the putative type V within class 2 (Makarova et al., 2015). The features of the putative type V loci, especially the domain architecture of Cpf1, suggest not only that type II and type V systems independently evolved through the association of different adaptation modules (*cas1*, *cas2*, and *cas4* genes) with different TnpB genes, but also that type V systems are functionally unique. The notion that Cpf1-carrying loci are bona fide CRISPR systems is further buttressed by the search of microbial genome sequences for similarity to the type V spacers that produced several significant hits to prophage genes—in particular, those from *Francisella* (Schunder et al., 2013). Given these observations and the prevalence of Cpf1-family proteins in diverse bacterial species, we sought to test the hypothesis that Cpf1-encoding CRISPR-Cas loci are biologically active and can mediate targeted DNA interference, one of the primary functions of CRISPR systems.

To simplify experimentation, we cloned the *Francisella novicida U112* Cpf1 (FnCpf1) locus (Figure 1A) into low-copy plasmids (pFnCpf1) to allow heterologous reconstitution in *Escherichia coli*. Typically, in currently characterized CRISPR-Cas systems, there are two requirements for DNA interference: (1) the target sequence has to match one of the spacers present in the respective CRISPR array, and (2) the target sequence complementary to the spacer (hereinafter protospacer) has to be flanked by the appropriate protospacer adjacent motif (PAM). Given the completely uncharacterized functionality of the FnCpf1 CRISPR locus, we adapted a previously described plasmid depletion assay (Jiang et al., 2013) to ascertain the activity of Cpf1 and identify the requirement for a PAM sequence and its respective location relative to the protospacer (5′ or 3′) (Figure 1B). We constructed two libraries of plasmids carrying a protospacer matching the first spacer in the FnCpf1 CRISPR array with the 5′ or 3′ 7 bp sequences randomized. Each plasmid library was transformed into *E. coli* that heterologously expressed the FnCpf1 locus or into a control *E. coli* strain carrying the empty vector. Using this assay, we determined the PAM sequence and location by identifying nucleotide motifs that are preferentially depleted in cells heterologously expressing the FnCpf1 locus. We found that the PAM for FnCpf1 is located upstream of the 5′ end of the displaced strand of the protospacer and has the sequence 5′-TTN (Figures 1C, 1D and S1). The 5′ location of the PAM is also observed in type I CRISPR systems, but not in type II systems, where Cas9 employs PAM sequences that are located on the 3′ end of the protospacer (Mojica et al., 2009; Garneau et al., 2010). Beyond the identification of the PAM, the results of the depletion assay clearly indicate that heterologously expressed Cpf1 loci are capable of efficient interference with plasmid DNA.

To further characterize the PAM requirements, we analyzed plasmid interference activity by transforming *cpf1*-locus-expressing cells with plasmids carrying protospacer 1 flanked by 5′-TTN PAMs. We found that all 5′-TTN PAMs were efficiently targeted (Figure 1E). In addition, 5′-CTA, but not 5′-TCA, was also efficiently targeted (Figure 1E), suggesting that the middle T is more critical for PAM recognition than the first T and that, in agreement with the sequence motifs depleted in the PAM discovery assay (Figure S1D), the PAM might be more relaxed than 5′-TTN.

The Cpf1-Associated CRISPR Array Is Processed Independent of TracrRNA

After showing that *cpf1*-based CRISPR loci are able to mediate robust DNA interference, we performed small RNA sequencing to determine the exact identity of the crRNA produced by these loci. By sequencing small RNAs extracted from a *Francisella novicida U112* culture, we found that the CRISPR array is processed into short mature crRNAs of 42–44 nt in length. Each mature crRNA begins with 19 nt of the direct repeat followed by 23–25 nt of the spacer sequence (Figure 2A). This crRNA arrangement contrasts with that of type II CRISPR-Cas systems in which the mature crRNA starts with 20–24 nt of spacer sequence followed by ~22 nt of direct repeat (Deltcheva et al., 2011; Chylinski et al., 2013). Unexpectedly, apart from the crRNAs, we did not observe any robustly expressed small transcripts near the *Francisella cpf1* locus that might correspond to tracrRNAs, which are associated with Cas9-based systems.

To confirm that no additional RNAs are required for crRNA maturation and DNA interference, we constructed an expression plasmid using synthetic promoters to drive the expression of *Francisella cpf1* (FnCpf1) and the CRISPR array (pFnCpf1_min). Small RNaseq of *E. coli* expressing this plasmid still showed robust processing of the CRISPR array into mature crRNA (Figure 2B), indicating that FnCpf1 and its CRISPR array are the only elements required from the FnCpf1 locus to achieve crRNA processing. Furthermore, *E. coli* expressing pFnCpf1_min as well as pFnCpf1_ΔCas, a plasmid with all of the *cas* genes removed but retaining native promoters driving the expression of FnCpf1 and the CRISPR array, also exhibited robust DNA interference, demonstrating that FnCpf1 and crRNA are sufficient for mediating DNA targeting (Figure 2C). By contrast, Cas9 requires both crRNA and tracrRNA to mediate targeted DNA interference (Deltcheva et al., 2011; Zhang et al., 2013).

Cpf1 Is a Single crRNA-Guided Endonuclease

The finding that FnCpf1 can mediate DNA interference with crRNA alone is highly surprising given that Cas9 recognizes crRNA through the duplex structure between crRNA and tracrRNA (Jinek et al., 2012; Nishimasu et al., 2014), as well as the 3′ secondary structure of the tracrRNA (Hsu et al., 2013; Nishimasu et al., 2014). To ensure that crRNA is indeed sufficient for forming an active complex with FnCpf1 and mediating RNA-guided DNA cleavage, we investigated whether FnCpf1 supplied only with crRNA can cleave target DNA in vitro. We purified

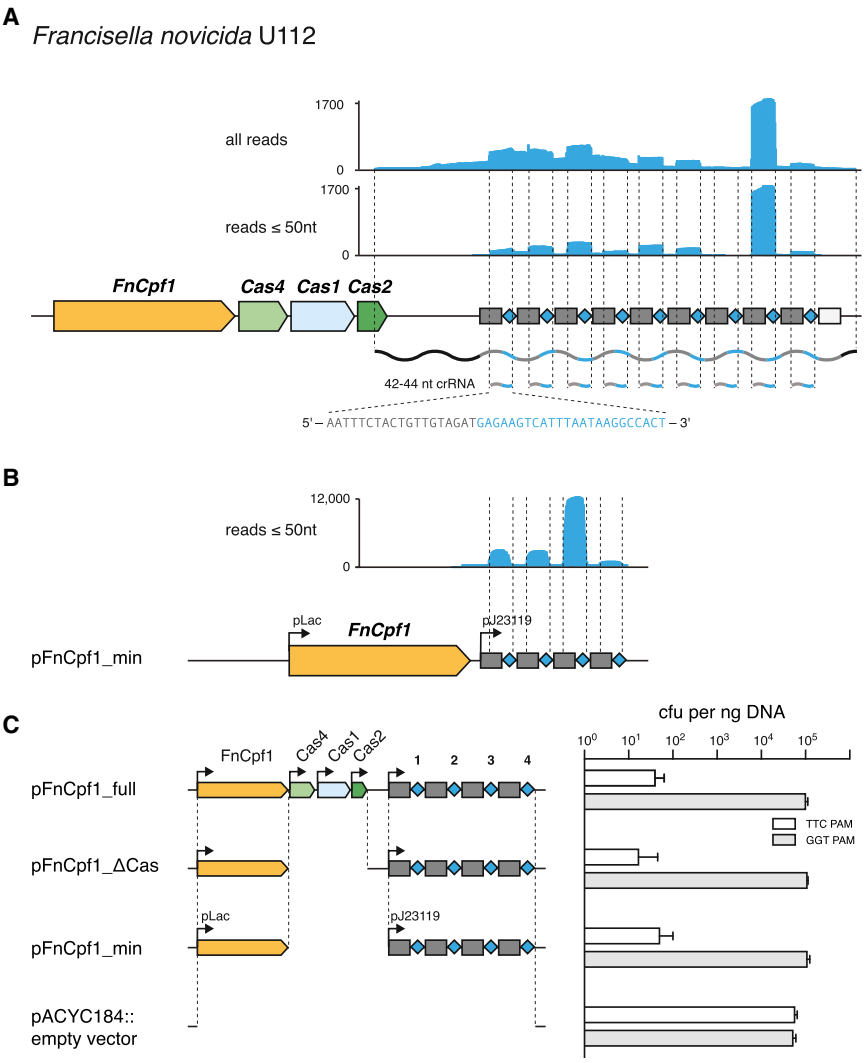


Figure 2. Heterologous Expression of FnCpf1 and CRISPR Array in *E. coli* Is Sufficient to Mediate Plasmid DNA Interference and crRNA Maturation

(A) Small RNA-seq of *Francisella novicida U112* reveals transcription and processing of the FnCpf1 CRISPR array. The mature crRNA begins with a 19-nt partial direct repeat followed by 23–25 nt of spacer sequence.

(B) Small RNA-seq of *E. coli* transformed with a plasmid-carrying synthetic promoter-driven FnCpf1 and CRISPR array shows crRNA processing independent of Cas genes and other sequence elements in the FnCpf1 locus.

(C) *E. coli* harboring different truncations of the FnCpf1 CRISPR locus shows that only FnCpf1 and the CRISPR array are required for plasmid DNA interference (n = 3; error bars show mean ± SEM).

FnCpf1 (Figure S2) and assayed its ability to cleave the same protospacer-1-containing plasmid used in the bacterial DNA interference experiments (Figure 3A). We found that FnCpf1 along with an in-vitro-transcribed mature crRNA-targeting protospacer 1 was able to efficiently cleave the target plasmid in a Mg^{2+}- and crRNA-dependent manner (Figure 3B). Moreover, FnCpf1 was able to cleave both supercoiled and linear target DNA (Figure 3C). These results clearly demonstrate the sufficiency of FnCpf1 and crRNA for RNA-guided DNA cleavage.

We also mapped the cleavage site of FnCpf1 using Sanger sequencing of the cleaved DNA ends. We found that FnCpf1-mediated cleavage results in a 5-nt 5′ overhang (Figures 3A, 3D, and S3A–S3D), which is different from the blunt cleavage product generated by Cas9 (Garneau et al., 2010; Jinek et al., 2012; Gasiunas et al., 2012). The staggered cleavage site of FnCpf1 is distant from the PAM: cleavage occurs after the 18th base on the non-targeted (+) strand and after the 23rd base on the targeted (–) strand (Figures 3A, 3D, and S3A–S3D). Using double-stranded oligo substrates with different PAM sequences, we also found that FnCpf1 requires the 5′-TTN PAM to be in a duplex form in order to cleave the target DNA (Figure 3E).

The RuvC-like Domain of Cpf1 Mediates RNA-Guided DNA Cleavage

The RuvC-like domain of Cpf1 retains all of the catalytic residues of this family of endonucleases (Figures 4A and S4) and is thus predicted to be an active nuclease. Therefore, we generated three mutants—FnCpf1(D917A), FnCpf1(E1006A), and FnCpf1(D1225A) (Figure 4A)—to test whether the conserved catalytic residues are essential for the nuclease activity of FnCpf1. We found that the D917A and E1006A mutations completely inactivated the DNA cleavage activity of FnCpf1, and D1255A significantly reduced nucleolytic activity (Figure 4B). These results are in contrast to the mutagenesis results for *Streptococcus pyogenes* Cas9 (SpCas9), where mutation of the RuvC (D10A) and HNH (N863A) nuclease domains converts SpCas9 into a DNA nickase (i.e., inactivation of each of the two nuclease domains abolished the cleavage of one of the DNA strands) (Jinek et al., 2012; Gasiunas et al., 2012) (Figure 4B). These findings suggest that the RuvC-like domain of FnCpf1 cleaves both strands of the target DNA, perhaps in a dimeric configuration. Interestingly, size-exclusion gel filtration of FnCpf1 shows that the protein is eluted at a size of ~300 kD, twice the molecular weight of a FnCpf1 monomer (Figure S2B).

Sequence and Structural Requirements for the Cpf1 crRNA

Compared with the guide RNA for Cas9, which has elaborate RNA secondary structure features that interact with Cas9 (Nishimasu et al., 2014), the guide RNA for FnCpf1 is notably simpler and only consists of a single stem loop in the direct repeat

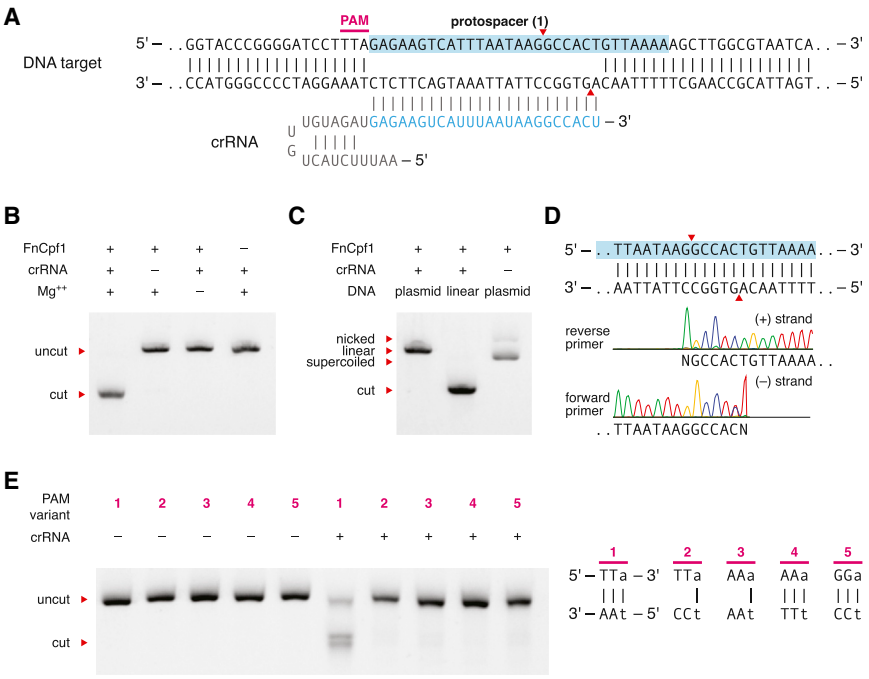


Figure 3. FnCpf1 Is Guided by crRNA to Cleave DNA In Vitro

(A) Schematic of the FnCpf1 crRNA-DNA-targeting complex. Cleavage sites are indicated by red arrows.

(B) FnCpf1 and crRNA alone mediated RNA-guided cleavage of target DNA in a crRNA- and Mg^{2+}-dependent manner.

(C) FnCpf1 cleaves both linear and supercoiled DNA.

(D) Sanger-sequencing traces from FnCpf1-digested target show staggered overhangs. The non-templated addition of an additional adenine, denoted as N, is an artifact of the polymerase used in sequencing (Clark, 1988). Reverse primer read represented as reverse complement to aid visualization. See also Figure S3.

(E) Dependency of cleavage on base-pairing at the 5′ PAM. FnCpf1 can only recognize the PAM in correctly Watson-Crick-paired DNA.

See also Figures S2 and S3.

sequence (Figure 3A). We explored the sequence and structural requirements of crRNA for mediating DNA cleavage with FnCpf1.

We first examined the length requirement for the guide sequence and found that FnCpf1 requires at least 16 nt of guide sequence to achieve detectable DNA cleavage and a minimum of 18 nt of guide sequence to achieve efficient DNA cleavage in vitro (Figure 5A). These requirements are similar to those demonstrated for SpCas9, in which a minimum of 16–17 nt of spacer sequence is required for DNA cleavage (Cencic et al., 2014; Fu et al., 2014). We also found that the seed region of the FnCpf1 guide RNA is approximately within the first 5 nt on the 5′ end of the spacer sequence (Figures 5B and S3E).

Next, we studied the effect of direct repeat mutations on the RNA-guided DNA cleavage activity. The direct repeat portion of mature crRNA is 19 nt long (Figure 2A). Truncation of the direct repeat revealed that at least 16, but optimally more than 17 nt, of the direct repeat is required for cleavage. Mutations in the stem loop that preserved the RNA duplex did not affect the cleavage activity, whereas mutations that disrupted the stem loop duplex structure completely abolished cleavage (Figure 5D). Finally, base substitutions in the loop region did not affect nuclease activity, whereas the uracil base immediately proceeding the spacer sequence could not be substituted (Figure 5E). Collectively, these results suggest that FnCpf1 recognizes the crRNA through a combination of sequence-specific and structural features of the stem loop.

Cpf1-Family Proteins from Diverse Bacteria Share Common crRNA Structures and PAMs

Based on our previous experience in harnessing Cas9 for genome editing in mammalian cells, only a small fraction of bacterial nucleases can function efficiently when heterologously expressed in mammalian cells (Cong et al., 2013; Ran et al., 2015). Therefore, in order to assess the feasibility of harnessing Cpf1 as a genome-editing tool, we exploited the diversity of Cpf1-family proteins available in the public sequences databases. A BLAST search of the WGS database at the NCBI revealed 46 non-redundant Cpf1-family proteins (Figure S5A), from which we chose 16 candidates that, based on our phylogenetic reconstruction (Figure S5A), represented the entire Cpf1 diversity (Figures 6A and S5). These Cpf1-family proteins span a range of lengths between ~1,200 and ~1,500 amino acids.

The direct repeat sequences for each of these Cpf1-family proteins show strong conservation in the 19 nt at the 3′ of the direct repeat, the portion of the repeat that is included in the processed crRNA (Figure 6B). The 5′ sequence of the direct repeat is much more diverse. Of the 16 Cpf1-family proteins chosen for analysis, three (2, *Lachnospiraceae bacterium MC2017*, Lb3Cpf1; 3, *Butyrivibrio proteoclasticus*, BpCpf1; and 6, *Smithella sp. SC_K08D17*, SsCpf1) were associated with direct repeat sequences that are notably divergent from the FnCpf1 direct repeat (Figure 6B). However, even these direct repeat sequences preserved stem-loop structures that were identical or nearly identical to the FnCpf1 direct repeat (Figure 6C).

Given the strong structural conservation of the direct repeats that are associated with many of the Cpf1-family proteins, we first tested whether the orthologous direct repeat sequences are able to support FnCpf1 nuclease activity in vitro. As expected, the direct repeats that contained conserved stem sequences were able to function interchangeably with FnCpf1. By contrast, the direct repeats from candidates 2 (Lb3Cpf1) and 6 (SsCpf1) were unable to support FnCpf1 cleavage activity (Figure 6D). The direct repeat from candidate 3 (BpCpf1) supported only a low level of FnCpf1 nuclease activity (Figure 6D), possibly due to the conservation of the 3′-most U.

Next, we applied the in vitro PAM identification assay (Figure S6A) to determine the PAM sequence for each Cpf1-family protein. We were able to identify the PAM sequence for seven

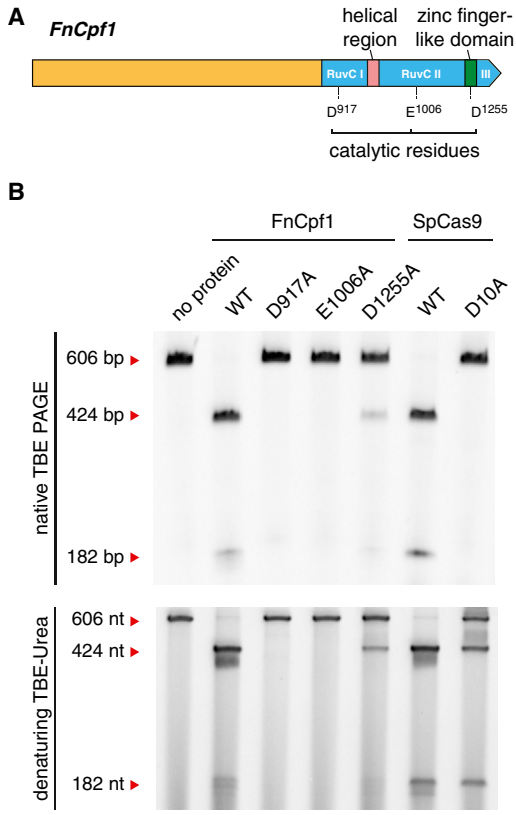


Figure 4. Catalytic Residues in the C-Terminal RuvC Domain of FnCpf1 Are Required for DNA Cleavage

(A) Domain structure of FnCpf1 with RuvC catalytic residues highlighted. The catalytic residues were identified based on sequence homology to *Thermus thermophilus* RuvC (PDB: 4EP5).

(B) Native TBE PAGE gel showing that mutation of the RuvC catalytic residues of FnCpf1 (D917A and E1006A) and mutation of the RuvC (D10A) catalytic residue of SpCas9 prevents double-stranded DNA cleavage. Denaturing TBE-Urea PAGE gel showing that mutation of the RuvC catalytic residues of FnCpf1 (D917A and E1006A) prevents DNA-nicking activity, whereas mutation of the RuvC (D10A) catalytic residue of SpCas9 results in nicking of the target site. See also Figure S4.

new Cpf1-family proteins (Figures 6E, S6B, and S6C), and the screen confirmed the PAM for FnCpf1 as 5′-TTN. The remaining eight tested Cpf1 proteins did not show efficient cleavage during in vitro reconstitution. The PAM sequences for the Cpf1-family proteins were predominantly T rich, only varying in the number of Ts constituting each PAM (Figures 6E, S6B, and S6C).

Cpf1 Can Be Harnessed to Facilitate Genome Editing in Human Cells

We tested each Cpf1-family protein for which we were able to identify a PAM for nuclease activity in mammalian cells. We codon optimized each of these genes and attached a C-terminal nuclear localization signal (NLS) for optimal expression and nuclear targeting in human cells (Figure 7A). To test the activity of each Cpf1-family protein, we selected a guide RNA target site within the *DNMT1* gene (Figure 7B). We first found that each of the Cpf1-family proteins along with its respective crRNA designed to target *DNMT1* was able to cleave a PCR amplicon of the *DNMT1* genomic region in vitro (Figure 7C). However, when tested in human embryonic kidney 293FT (HEK293FT) cells, only two out of the eight Cpf1-family proteins (7, AsCpf1 and 13, LbCpf1) exhibited detectable levels of nuclease-induced indels (Figures 7C and 7D). This result is consistent with previous experiments with Cas9 in which only a small number of Cas9 orthologs were successfully harnessed for genome editing in mammalian cells (Ran et al., 2015).

We further tested each Cpf1-family protein with additional genomic targets and found that AsCpf1 and LbCpf1 consistently mediated robust genome editing in HEK293FT cells, whereas the remaining Cpf1 proteins showed either no detectable activity or only sporadic activity (Figures 7E and S7) despite robust expression (Figure S6D). The only Cpf1 candidate that expressed poorly was PdCpf1 (Figure S6D). When compared to Cas9, AsCpf1 and LbCpf1 mediated comparable levels of indel formation (Figure 7E). Additionally, we used in vitro cleavage followed by Sanger sequencing of the cleaved DNA ends and found that 7, AsCpf1 and 13, LbCpf1 also generated staggered cleavage sites (Figures S6E and S6F, respectively).

DISCUSSION

In this work, we characterize Cpf1-containing class 2 CRISPR systems, classified as type V, and show that its effector protein, Cpf1, is a single RNA-guided endonuclease. Cpf1 substantially differs from Cas9—to date, the only other experimentally characterized class 2 effector—in terms of structure and function and might provide important advantages for genome-editing applications. Specifically, Cpf1 contains a single identified nuclease domain, in contrast to the two nuclease domains present in Cas9. The results presented here show that, in FnCpf1, inactivation of RuvC-like domain abolishes cleavage of both DNA strands. Conceivably, FnCpf1 forms a homodimer (Figure S2B), with the RuvC-like domains of each of the two subunits cleaving one DNA strand. However, we cannot rule out that FnCpf1 contains a second yet-to-be-identified nuclease domain. Structural characterization of Cpf1-RNA-DNA complexes will allow testing of these hypotheses and elucidation of the cleavage mechanism.

Perhaps the most notable feature of Cpf1 is that it is a single crRNA-guided endonuclease. Unlike Cas9, which requires tracrRNA to process crRNA arrays and both crRNA and tracrRNA to mediate interference (Deltcheva et al., 2011), Cpf1 processes crRNA arrays independent of tracrRNA, and Cpf1-crRNA complexes alone cleave target DNA molecules, without the requirement for any additional RNA species. This feature could simplify the design and delivery of genome-editing tools. For example, the shorter (~42 nt) crRNA employed by Cpf1 has practical advantages over the long (~100 nt) guide RNA in Cas9-based systems because shorter RNA oligos are significantly easier and cheaper to synthesize. In addition, these findings raise more fundamental questions regarding the guide processing mechanism of the type V CRISPR-Cas systems. In the case of type II, processing of the pre-crRNA is catalyzed by the bacterial RNase III, which recognizes the long duplex formed by the tracrRNA and the complementary portion of the direct repeat (Deltcheva et al., 2011). Such long duplexes

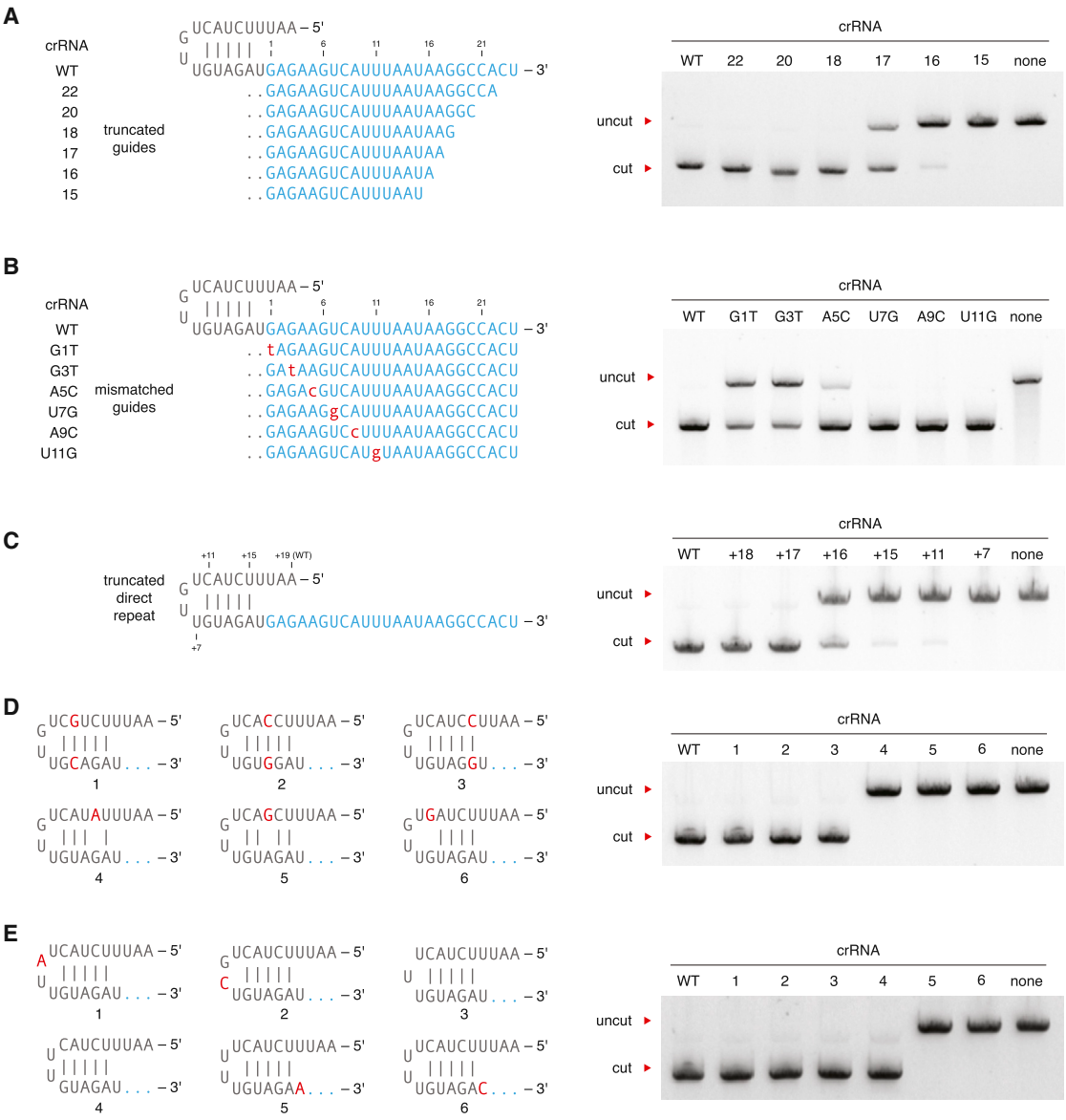


Figure 5. crRNA Requirements for FnCpf1 Nuclease Activity In Vitro

(A) Effect of spacer length on FnCpf1 cleavage activity.

(B) Effect of crRNA-target DNA mismatch on FnCpf1 cleavage activity. See also Figure S3E.

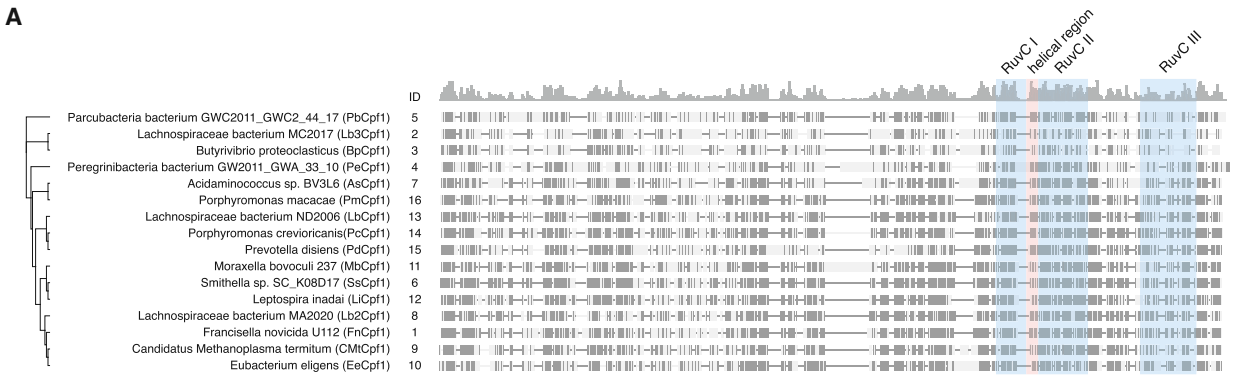
(C) Effect of direct repeat length on FnCpf1 cleavage activity.

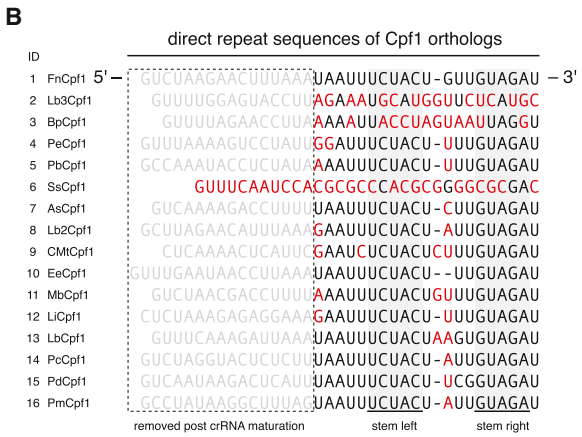
(D) FnCpf1 cleavage activity depends on secondary structure in the stem of the direct repeat RNA structure.

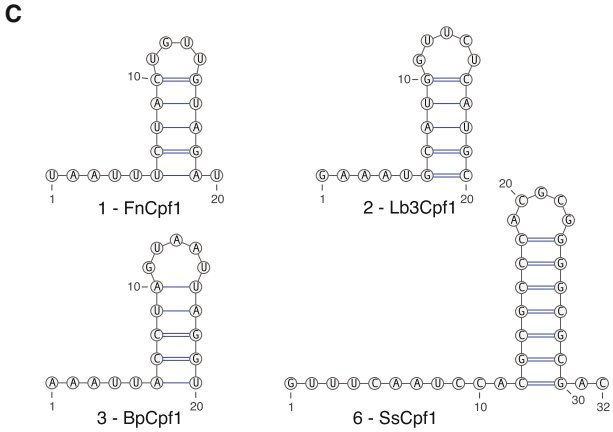
(E) FnCpf1 cleavage activity is unaffected by loop mutations but is sensitive to mutation in the 3′-most base of the direct repeat. See also Figure S4.

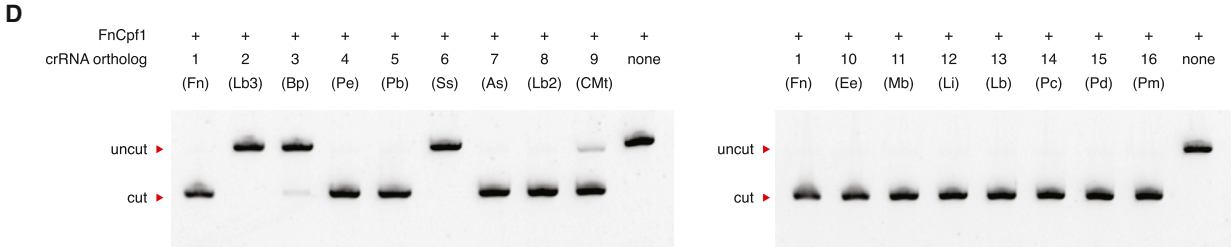
are not present in the pre-crRNA of type V systems, making it unlikely that RNase III is responsible for processing. Further experiments aimed at elucidating the processing mechanism of type V systems will shed light on the functional diversity of different CRISPR-Cas systems.

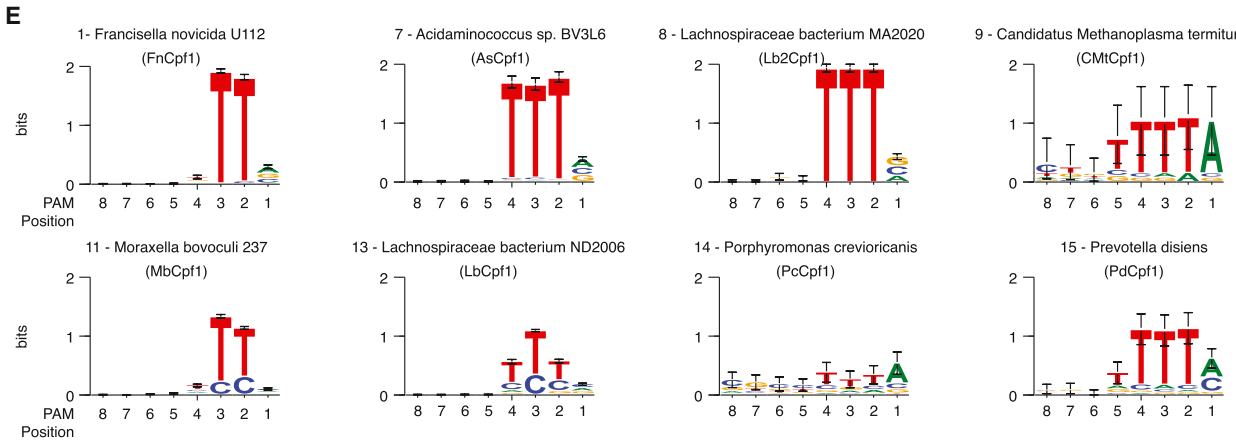
Cpf1 generates a staggered cut with a 5′ overhang, in contrast to the blunt ends generated by Cas9 (Garneau et al., 2010; Jinek et al., 2012; Gasiunas et al., 2012). This structure of the cleavage product could be particularly advantageous for facilitating non-homologous end joining (NHEJ)-based gene insertion into the mammalian genome (Maresca et al., 2013). Being able to program the exact sequence of a sticky end would allow researchers to design the DNA insert so that it integrates into the genome in the proper orientation. Specifically, in non-dividing cells, in which genome editing via homology-directed repair (HDR) mechanisms is especially challenging (Chan et al., 2011), Cpf1 could provide an effective way to precisely introduce DNA into the genome via non-HDR mechanisms.











(legend on next page)

Another potentially useful feature of Cpf1 that might aid the introduction of new DNA sequences is that Cpf1 cleaves target DNA at the distal end of the protospacer, far away from the seed region. Therefore, Cpf1-induced indels will be located far from the target site, which is thus preserved for subsequent rounds of Cpf1 cleavage. With Cas9, any indel resulting from the dominant NHEJ repair pathway will disrupt the target site, effectively eliminating the possibility of inserting new DNA at that site in that particular cell. In the case of Cpf1, it appears possible that, if the first round of targeting results in an indel, a subsequent round of targeting could yet be repaired via HDR. Future exploration of these and other strategies using Cpf1 and other class 2 effectors is expected to bring solutions for some of the biggest challenges facing genome editing.

The T-rich PAMs of the Cpf1-family also allow for applications in genome editing in organisms with particularly AT-rich genomes, such as *Plasmodium falciparum* (Gardner et al., 2002) or areas of interest with AT enrichment, such as scaffold/matrix attachment regions. To date, all characterized mammalian genome-editing proteins require the presence of at least one G (Hsu et al., 2014; Jiang et al., 2015), so the T- and T/C-dependent PAMs of Cpf1-family proteins expand the targeting range of RNA-guided genome editing nucleases.

The natural diversity of CRISPR systems provides a wealth of opportunities for understanding the origin and evolution of prokaryotic adaptive immunity, as well as for harnessing potentially transformative biotechnological tools. There is little doubt that, beyond the already classified and characterized diversity of the CRISPR-Cas types, there are additional systems with distinctive characteristics that await exploration and could further enhance genome editing and other areas of biotechnology as well as shed further light on the evolution of these defense systems.

EXPERIMENTAL PROCEDURES

Generation of Heterologous Plasmids

To generate the FnCpf1 locus for heterologous expression, genomic DNA from *Francisella novicida* (generous gift from Wayne Conlan) was PCR amplified using Herculase II polymerase (Agilent Technologies) and cloned into pACYC-184 using Gibson cloning (New England Biolabs). Cells harboring plasmids were made competent using the Z-competent kit (Zymo). Sequences of all bacterial expression plasmids can be found in Table S1.

Bacterial RNA Sequencing

RNA was isolated from stationary-phase bacteria by first resuspending *F. novicida* (generous gift from David Weiss) or *E. coli* in TRIzol and then homogenizing the bacteria with zirconia/silica beads (BioSpec Products) in a BeadBeater (BioSpec Products) for three 1-min cycles. Total RNA was purified from homogenized samples with the Direct-Zol RNA miniprep protocol (Zymo), DNase treated with TURBO DNase (Life Technologies), and 3′ dephosphorylated with T4 Polynucleotide Kinase (New England Biolabs). rRNA was removed with the bacterial Ribo-Zero rRNA removal kit (Illumina). RNA libraries were prepared from rRNA-depleted RNA using NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs) and size selected using the Pippin Prep (Sage Science).

For heterologous *E. coli* expression of the FnCpf1 locus, RNA-sequencing libraries were prepared from rRNA-depleted RNA using a derivative of the previously described CRISPR RNA-sequencing method (Heidrich et al., 2015). In brief, transcripts were poly-A tailed with *E. coli* Poly(A) Polymerase (New England Biolabs), ligated with 5′ RNA adapters using T4 RNA Ligase 1 (ssRNA Ligase) High Concentration (New England Biolabs), and reverse transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent Technologies). cDNA was PCR amplified with barcoded primers using Herculase II polymerase (Agilent Technologies).

RNA-Sequencing Analysis

The prepared cDNA libraries were sequenced on a MiSeq (Illumina). Reads from each sample were identified on the basis of their associated barcode and aligned to the appropriate RefSeq reference genome using BWA (Li and Durbin, 2009). Paired-end alignments were used to extract entire transcript sequences using Picard tools (http://broadinstitute.github.io/picard), and these sequences were analyzed using Geneious 8.1.5 (Biomatters).

In Vivo FnCpf1 PAM Screen

Randomized PAM plasmid libraries were constructed using synthesized oligonucleotides (IDT) consisting of eight or seven randomized nucleotides either upstream or downstream, respectively, of the FnCpf1 spacer 1. The randomized ssDNA oligos (Table S1) were made double stranded by annealing to a short primer and using the large Klenow fragment (New England Biolabs) for second-strand synthesis. The dsDNA product was assembled into a linearized pUC19 using Gibson cloning (New England Biolabs). Competent Stbl3 *E. coli* (Invitrogen) were transformed with the cloned products, and $>10^7$ cells were collected and pooled. Plasmid DNA was harvested using a Maxi-prep kit (QIAGEN). We transformed 30 ng of the pooled library into *E. coli* cells carrying the FnCpf1 locus or pACYC184 control. After transformation, cells were plated on ampicillin. After 16 hr of growth, >4E6 cells were harvested and plasmid DNA was extracted using a Maxi-prep kit (QIAGEN). The target PAM region was amplified and sequenced using a MiSeq (Illumina) with a single-end 150 cycle kit.

Computational PAM Discovery Pipeline

PAM regions were extracted, counted, and normalized to total reads for each sample. For a given PAM, enrichment was measured as the log ratio compared to pACYC184 control, with a 0.01 psuedocount adjustment. PAMs above a 3.5 enrichment threshold were collected and used to generate sequence logos (Crooks et al., 2004).

PAM Validation

Sequences corresponding to both PAMs and non-PAMs were cloned into digested pUC19 and ligated with T4 ligase (Enzymatics). Competent *E. coli* with either the FnCpf1 locus plasmid or pACYC184 control plasmid were transformed with 20 ng of PAM plasmid and plated on LB agar plates supplemented with ampicillin and chloramphenicol. Colonies were counted after 18 hr.

Figure 6. Analysis of Cpf1-Family Protein Diversity and Function

(A) Phylogenetic tree of 16 Cpf1 orthologs selected for functional analysis. Conserved sequences are shown in dark gray. The RuvC domain, helical region, and zinc finger are highlighted.

(B) Alignment of direct repeats from the 16 Cpf1-family proteins. Sequences that are removed post crRNA maturation are colored gray. Non-conserved bases are colored red. The stem duplex is highlighted in gray.

(C) RNAfold (Lorenz et al., 2011) prediction of the direct repeat sequence in the mature crRNA. Predictions for FnCpf1 along with three diverged type V loci are shown.

(D) Type V crRNAs from different bacteria with similar direct repeat sequences are able to function with FnCpf1 to mediate target DNA cleavage.

(E) PAM sequences for eight Cpf1-family proteins identified using in vitro cleavage of a plasmid library containing randomized PAMs flanking the protospacer. See also Figures S5 and S6.

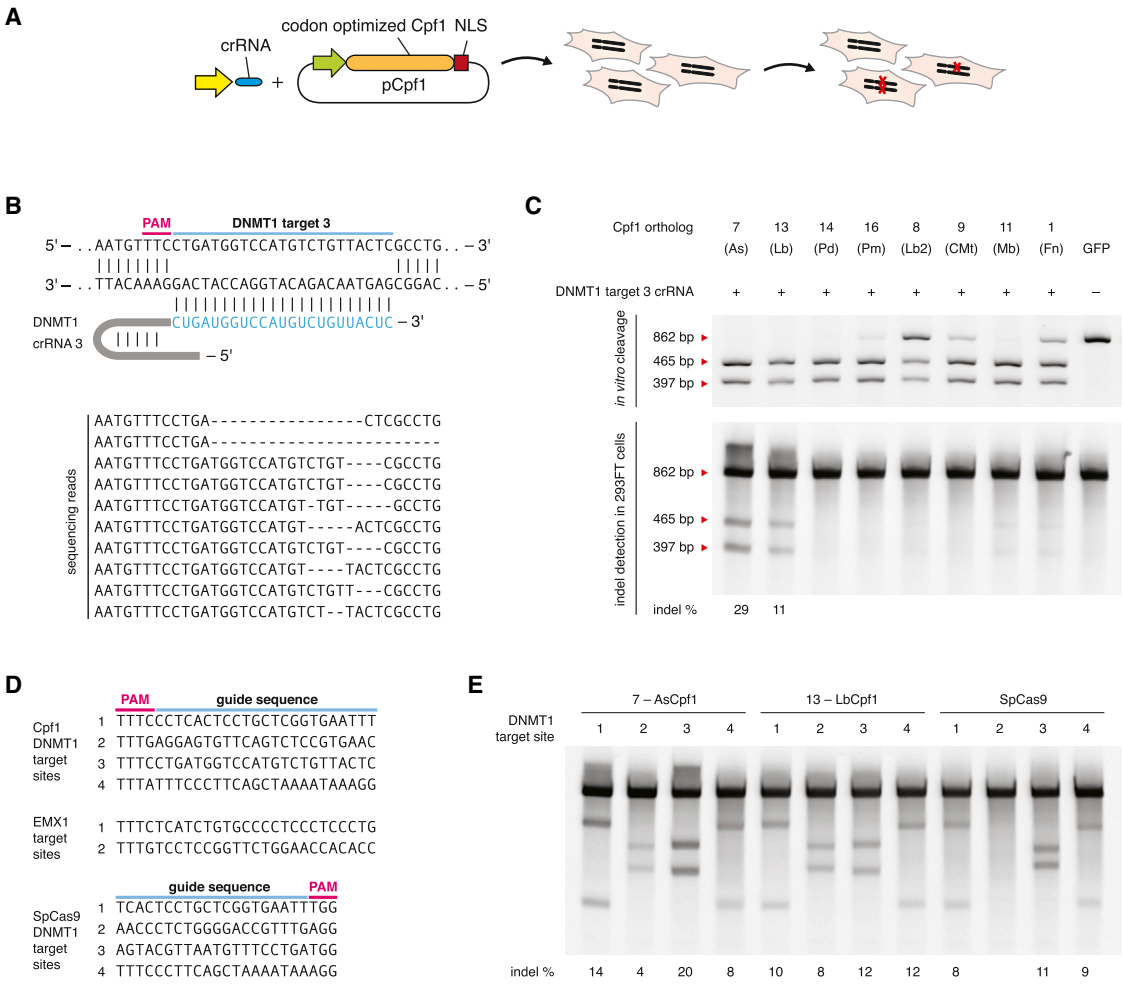


Figure 7. Cpf1 Mediates Robust Genome Editing in Human Cell Lines

(A) Eight Cpf1-family proteins were individually expressed in HEK293FT cells using CMV-driven expression vectors. The corresponding crRNA was expressed using a PCR fragment containing a U6 promoter fused to the crRNA sequence. Transfected cells were analyzed using either Surveyor nuclease assay or targeted deep sequencing.

(B) Schematic showing the sequence of DNMT1-targeting crRNA 3. Sequencing reads show representative indels.

(C) Comparison of in vitro and in vivo cleavage activity. The DNMT1 target region was PCR amplified, and the genomic fragment was used to test Cpf1-mediated cleavage. All eight Cpf1-family proteins showed DNA cleavage in vitro (top), but only candidates 7, AsCpf1 and 13, Lb3Cpf1 facilitated robust indel formation in human cells.

(D) Cpf1 and SpCas9 target sequences in the human DNMT1 and EMX1 loci.

(E) Comparison of Cpf1 and SpCas9 genome-editing efficiency. Target sites correspond to sequences shown in Figure 7D.

See also Figure S7.

Synthesis of crRNAs and sgRNAs

All crRNAs and sgRNAs used in biochemical reactions were synthesized using the HiScribe T7 High Yield RNA Synthesis Kit (NEB). ssDNA oligos (Table S2) corresponding to the reverse complement of the target RNA sequence were synthesized from IDT and annealed to a short T7 priming sequence. T7 transcription was performed for 4 hr, and then RNA was purified using the MEGAclear Transcription Clean-Up Kit (Ambion).

Purification of Cpf1 Protein

FnCpf1 protein was cloned into a bacterial expression vector (6-His-MBP-TEV-Cpf1, a pET based vector generously provided by Doug Daniels). Two liters of Terrific Broth growth media with 100 μg/ml ampicillin were inoculated with 10 ml overnight culture Rosetta (DE3) pLyseS (EMD Millipore) cells containing the Cpf1 expression construct. Growth media plus inoculant was grown at 37°C until the cell density reached 0.2 OD600, then the temperature was decreased to 21°C. Growth was continued until OD600 reached 0.6 when a final concentration of 500 μM IPTG was added to induce MBP-Cpf1 expression. The culture was induced for 14–18 hr before harvesting cells and freezing at −80°C until purification.

Cell paste was resuspended in 200 ml of Lysis Buffer (50 mM HEPES [pH 7], 2M NaCl, 5 mM $MgCl_2$, 20 mM imidazole) supplemented with protease inhibitors (Roche cOmplete, EDTA-free) and lysozyme (Sigma). Once homogenized, cells were lysed by sonication (Branson Sonifier 450) and then centrifuged at 10,000 × *g* for 1 hr to clear the lysate. The lysate was filtered through 0.22 micron filters (Millipore, Stericup) and applied to a nickel column (HisTrap FF, 5 ml), washed, and then eluted with a gradient of imidazole. Fractions containing protein of the expected size were pooled, TEV protease (Sigma) was added, and the sample was dialyzed overnight into TEV buffer

(500 mM NaCl, 50 mM HEPES [pH 7], 5 mM MgCl, 2 mM DTT). After dialysis, TEV cleavage was confirmed by SDS-PAGE, and the sample was concentrated to 500 μl prior to loading on a gel filtration column (HiLoad 16/600 Superdex 200) via FPLC (AKTA Pure). Fractions from gel filtration were analyzed by SDS-PAGE; fractions containing Cpf1 were pooled and concentrated to 200 μl and either used directly for biochemical assays or frozen at −80°C for storage. Gel filtration standards were run on the same column equilibrated in 2M NaCl, HEPES (pH 7.0) to calculate the approximate size of FnCpf1.

Generation of Cpf1 Protein Lysate

Cpf1 proteins codon optimized for human expression were synthesized with a C-terminal nuclear localization tag and cloned into the pcDNA3.1 expression plasmid by Genscript (Table S1). 2,000 ng of Cpf1 expression plasmids were transfected into 6-well plates of HEK293FT cells at 90% confluency using Lipofectamine 2000 reagent (Life Technologies). 48 hr later, cells were harvested by washing once with DPBS (Life Technologies) and scraping in lysis buffer (20 mM HEPES [pH 7.5], 100 mM KCl, 5mM $MgCl_2$, 1 mM DTT, 5% glycerol, 0.1% Triton X-100, 1X cOmplete Protease Inhibitor Cocktail Tablets [Roche]). Lysate was sonicated for 10 min in a Biorupter sonicator (Diagenode) and then centrifuged. Supernatant was frozen for subsequent use in in vitro cleavage assays.

In Vitro Cleavage Assay

Cleavage in vitro was performed either with purified protein (25 nM) or mammalian lysate with protein at 37°C in cleavage buffer (NEBuffer 3, 5 mM DTT) for 20 min. The cleavage reaction used 500 ng of synthesized crRNA or sgRNA and 200 ng of target DNA. Target DNA involved either protospacers cloned into pUC19 or PCR amplicons of gene regions from genomic DNA isolated from HEK293 cells. Reactions were cleaned up using PCR purification columns (QIAGEN) and were run on 2% agarose E-gels (Life Technologies). For native and denaturing gels to analyze cleavage by nuclease mutants, cleaned-up reactions were run on TBE 6% polyacrylamide or TBE-Urea 6% polyacrylamide gels (Life Technologies).

In Vitro Cpf1-Family Protein PAM Screen

In vitro cleavage reactions with Cpf1-family proteins were run on 2% agarose E-gels (Life Technologies). Bands corresponding to un-cleaved target were gel extracted using QIAquick Gel Extraction Kit (QIAGEN), and the target PAM region was amplified and sequenced using a MiSeq (Illumina) with a single-end 150 cycle kit. Sequencing results were entered into the PAM discovery pipeline.

Western Blot Analysis

Cells were lysed in 1×RIPA buffer (Cell Signaling Technology) supplemented with protease inhibitor cocktail (Roche). Equal volumes of cell lysate were run on BOLT 4%–12% Bis-Tris gradient gels (Invitrogen) and transferred to PVDF membranes (Millipore). Non-specific antigen binding was blocked with TBS-T (50 mM Tris, 150 mM NaCl and 0.05% Tween-20) with 5% BLOT-QuickBlocker Reagent (Millipore) for 1 hr. Membranes were incubated with primary antibodies (anti-HA-tag [Cell Signaling Technology C29F4] or HRP-conjugated GAPDH [Cell Signaling Technology 14C10]) for 1 hr in TBS-T with 1% BLOT-QuickBlocker. Membranes were washed for three 10 min washes and anti-HA-tag membranes were further incubated with anti-rabbit antibody (Cell Signaling Technology 7074) for 1 hr followed by six 10 min washes in TBS-T. Proteins were visualized with West Pico Chemiluminescent Substrate (Life Technology) and imaged using the ChemiDoc MP Imaging System (Bio-Rad) and processed with ImageLab software (Bio-Rad).

SURVEYOR Nuclease Assay for Genome Modification

PCR amplicons comprised of a U6 promoter driving expression of the crRNA sequence were generated using Herculase II (Agilent Technologies) and appropriate U6 reverse primers (Table S2). 400 ng of Cpf1 expression plasmids and 100 ng of the U6::crRNA expression cassettes were transfected into 24-well plates of HEK293FT cells at 75%–90% confluency using Lipofectamine 2000 (Life Technologies).

Cells were incubated at 37°C for 72 hr post-transfection before genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicenter) following the manufacturer's protocol. The genomic region flanking the CRISPR target site for each gene was PCR amplified, and products were purified using QiaQuick Spin Column (QIAGEN) following the manufacturer's protocol. 200–500 ng total of the purified PCR products were mixed with 1 μl 10 × Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 10 μl and were subjected to a re-annealing process to enable heteroduplex formation: 95°C for 10 min, 95°C to 85°C ramping at −2°C/s, 85°C to 25°C at −0.25°C/s, and 25°C hold for 1 min. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Integrated DNA Technologies) following the manufacturer's recommended protocol and analyzed on 4%–20% Novex TBE polyacrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 10 min and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, $100 \times (1 - \mathrm{sqrt}(1 - (b + c)/(a + b + c)))$, where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Deep Sequencing to Characterize Cpf1 Indel Patterns in 293FT Cells

HEK293FT cells were transfected and harvested as described for assessing activity of Cpf1 cleavage. The genomic-region-flanking DNMT1 targets were amplified using a two-round PCR region to add Illumina P5 adapters as well as unique sample-specific barcodes to the target amplicons. PCR products were run on 2% E-gel (Invitrogen) and gel extracted using QiaQuick Spin Column (QIAGEN) as per the manufacturer's recommended protocol. Samples were pooled and quantified by Qubit 2.0 Fluorometer (Life Technologies). The prepared cDNA libraries were sequenced on a MiSeq with a single-end 300 cycle kit (Illumina). Indels were mapped using a Python implementation of the Geneious 6.0.3 Read Mapper.

Computational Analysis of Cpf1 loci

PSI-BLAST program (Altschul et al., 1997) was used to identify Cpf1 homologs in the NCBI NR database using several known Cpf1 sequences as queries with the Cpf1 with the E-value cut-off of 0.01 and low-complexity filtering and composition-based statistics turned off. The TBLASTN program with the E-value cut-off of 0.01 and low-complexity filtering turned off was used to search the NCBI WGS database using the Cpf1 profile (Makarova et al., 2015) as the query. Results of all searches were combined (Table S3). The HHpred program was used with default parameters (Söding et al., 2006) to identify remote sequence similarity using a subset of representative Cpf1 sequences queries. Multiple sequence alignments were constructed using MUSCLE (Edgar, 2004) with manual correction based on pairwise alignments obtained using PSI-BLAST and HHpred programs. Phylogenetic analysis was performed using the FastTree program with the WAG evolutionary model and the discrete gamma model with 20 rate categories (Price et al., 2010). Protein secondary structure was predicted using Jpred 4 (Drozdetskiy et al., 2015). CRISPR repeats were identified using PILER-CR (Edgar, 2007) and CRISPRfinder (Grissa et al., 2007).

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.038.

AUTHOR CONTRIBUTIONS

F.Z., B.Z., and J.S.G. conceived this study. B.Z., J.S.G., O.O.A., and F.Z. designed the experiments. B.Z., J.S.G., O.O.A., J.J., S.E.V., P.E., and F.Z. performed the experiments and analyzed the data. I.M.S. conducted FnCpf1 purification. A.R. assisted with RNA sequencing and analysis. K.S.M., E.V.K., and F.Z. performed the computational sequence analysis. F.Z., E.V.K., B.Z., J.S.G., O.O.A., K.S.M., and J.v.d.O. wrote the manuscript, which was read and approved by all authors.

ACKNOWLEDGMENTS

We would like to thank R. Macrae for critical reading of the manuscript. We would like to thank Wayne Conlan and David S. Weiss for generously providing *F. novicida* genomic DNA and cell pellets, respectively, Doug Daniels for kindly providing us with the bacterial expression vector, and Sergei Shmakov and Yuri Wolf for help with sequence analysis. E.V.K. and K.S.M are supported by the intramural program of the US Department of Health and Human Services (to the National Library of Medicine). J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. F.Z. is supported by the NIMH (5DP1-MH100706), the Poitras Center, Vallee, Simons, Paul G. Allen, and New York Stem Cell Foundations, David R. Cheng, Tom Harriman, and Bob Metcalfe. A patent application has been filed related to this work, and the authors plan to make the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab website (www.genome-engineering.org). F.Z. is a founder of Editas Medicine and a scientific advisor for Editas Medicine and Horizon Discovery.

Received: August 28, 2015
Revised: September 15, 2015
Accepted: September 17, 2015
Published: September 25, 2015

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol. Cell *54*, 234–244.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Cencic, R., Miura, H., Malina, A., Robert, F., Ethier, S., Schmeing, T.M., Dostie, J., and Pelletier, J. (2014). Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. PLoS ONE *9*, e109213.

Chan, F., Hauswirth, W.W., Wensel, T.G., and Wilson, J.H. (2011). Efficient mutagenesis of the rhodopsin gene in rod photoreceptor neurons in mice. Nucleic Acids Res. *39*, 5955–5966.

Charpentier, E., Richter, H., van der Oost, J., and White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol. Rev. *39*, 428–441.

Chylinski, K., Le Rhun, A., and Charpentier, E. (2013). The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol. *10*, 726–737.

Clark, J.M. (1988). Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. Nucleic Acids Res. *16*, 9677–9686.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein secondary structure prediction server. Nucleic Acids Res. *43*, W389–W394.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat. Biotechnol. *32*, 279–284.

Gardner, M.J., Shallom, S.J., Carlton, J.M., Salzberg, S.L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., et al. (2002). Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature *419*, 531–534.

Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. USA *109*, E2579–E2586.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. *35*, W52–W57.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell *139*, 945–956.

Heidrich, N., Dugar, G., Vogel, J., and Sharma, C. (2015). Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. In CRISPR, M. Lundgren, E. Charpentier, and P.C. Fineran, eds. (Springer New York), pp. 1–21.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.

Jackson, R.N., Golden, S.M., van Erp, P.B., Carter, J., Westra, E.R., Brouns, S.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science *345*, 1473–1479.

Jiang, W., and Marraffini, L.A. (2015). CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. Annu. Rev. Microbiol. Published online July 22, 2015. http://dx.doi.org/10.1146/annurev-micro-091014-104441.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Makarova, K.S., and Koonin, E.V. (2015). Annotation and classification of CRISPR-Cas systems. Methods Mol. Biol. *1311*, 47–75.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O., Costa, F.S.S., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). Updated evolutionary classification of CRISPR-Cas systems and cas genes. Nat. Rev.

Microbiol. Published online September 29, 2015. http://dx.doi.org/10.1038/nrmicro3534.

Maresca, M., Lin, V.G., Guo, N., and Yang, Y. (2013). Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. Genome Res. *23*, 539–546.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843–1845.

Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S., Héroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479–1484.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935–949.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186–191.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. *39*, 9275–9282.

Schunder, E., Rydzewski, K., Grunow, R., and Heuner, K. (2013). First indication for a functional CRISPR/Cas system in Francisella tularensis. Int. J. Med. Microbiol. *303*, 51–60.

Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. EMBO J. *32*, 385–394.

Söding, J., Remmert, M., Biegert, A., and Lupas, A.N. (2006). HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res. *34*, W374–W378.

Sorek, R., Lawrence, C.M., and Wiedenheft, B. (2013). CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu. Rev. Biochem. *82*, 237–266.

Vestergaard, G., Garrett, R.A., and Shah, S.A. (2014). CRISPR adaptive immune systems of Archaea. RNA Biol. *11*, 156–167.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol. Cell *50*, 488–503.

Cell

Correction

Morgan's Legacy: Fruit Flies and the Functional Annotation of Conserved Genes

Hugo J. Bellen* and Shinya Yamamoto
*Correspondence: hbellen@bcm.edu
http://dx.doi.org/10.1016/j.cell.2015.10.021

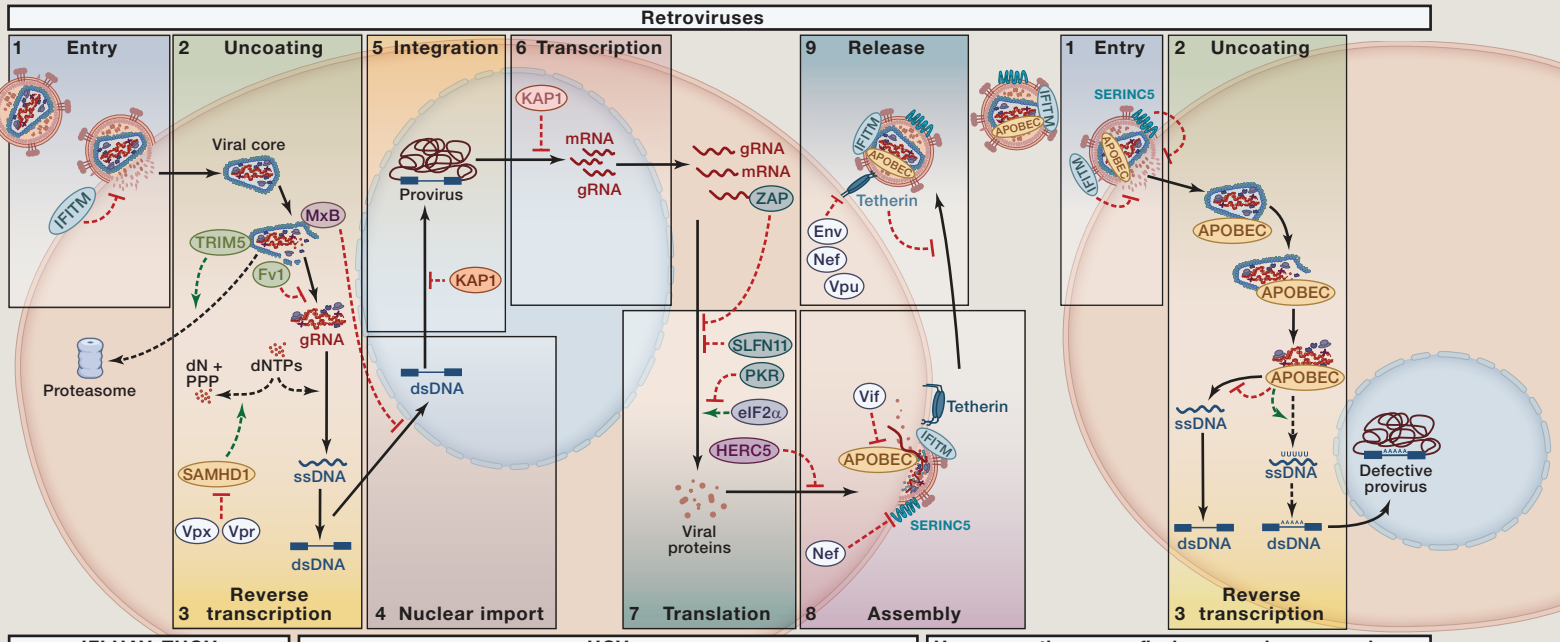
(Cell *163*, 12–14; September 24, 2015)

Due to a production error, a label in Figure 1 of this BenchMarks article was incorrect. The DNA element between UAS and PolyA should have read "cDNA," not "GAL4." The figure has been corrected online.

SnapShot: Antiviral Restriction Factors

Silvia F. Kluge, Daniel Sauter, and Frank Kirchhoff
Institute of Molecular Virology, Ulm University Medical Center, 89081 Ulm, Germany

Cell



	Restriction factor	Virus(es) targeted	Mechanism(s) of restriction	Viral antagonist(s) or evasion mechanism(s)	IFN-inducible	Positive selection
1	IFITM family	Retro-, orthomyxo-, flavi-, filo-, corona-, rhabdo-, bunya-, reoviruses	Inhibits membrane fusion, modification of lipid components, or membrane fluidity	None known	Some	Some
	SERINC3, 5	Retroviruses (HIV, SIV, MLV, EIAV)	Reduces membrane fusion	Nef (HIV, SIV), Glyco-Gag (MLV), S2 (EIAV)	N	N
	CH25H	Flavi-, retro-, filo-, bunya-, rhabdo-, herpesviruses	Inhibits membrane fusion by generating 25-hydroxycholesterol	None known	Y	Unknown
2	TRIM5α, TRIM-Cyp	Retroviruses (HIV, SIV, MLV, EIAV)	Accelerates uncoating, thereby inhibiting reverse transcription	Capsid mutation	Y	Y
	Fv1	MLV	Targets the viral capsid protein and interferes with uncoating	Capsid mutation	N	Y
3	APOBEC3 family	Hepadna-, retroviruses	Induces hypermutation by deamination, inhibits reverse transcription of HIV by binding to RNA and suppressing tRNA3Lys priming	Vif (lentiviruses), Bet (spumaviruses), Gag (gammaretroviruses)	Some	Some
	SAMHD1	Retroviruses	Hydrolyzes cellular dTNP and degrades viral RNA	Vpx (HIV-2, some SIV), Vpr (some SIV)	Y	Y
4	MxA	IFLUAV, THOV	Inhibits vRNP nuclear import	Nucleoprotein mutations (pandemic IFLUAV)	Y	Y
	MxB	HIV, SIV	Prevents integration of proviral DNA by inhibiting uncoating, nuclear uptake, and/or integrity/stability of the PIC	Capsid mutation	Y	Y
5	KAP1/TRIM28	HIV-1	Induces deacetylation of HIV integrase	None known	N	N
6	KAP1/TRIM28	Herpes-, retroviruses	Silences transcription and induces latency	vPK (KSHV)	N	N
	Viperin/RSAD2	HCV, DENV	Inhibits formation of the HCV replicon complex by sequestration of hVAP-33 and interaction with NS5A, interacts with NS3 (DENV)	None known	Y	Y
	CH25H	HCV	Inhibits membranous web formation and NS5A dimerization	None known	Y	Unknown
	IFI16	HPV, HCMV, HSV1	Accumulates on the viral genome and prevents association of transcriptional activators, induces heterochromatin formation	pUL97, pUL83 (HCMV)	Y	Y
	MxA	Bunyaviruses (LACV, RVFV, BUNV)	Sequesters newly synthesized viral N protein into perinuclear complexes	None known	Y	Y
	RNaseL (+OAS1)	Picorna-, flavi-, toga-, corona-, reo-, pox-, orthomyxo-, paramyxo-, herpes-, retro-, rhabdo-, hepadna-, polyomaviruses	Degrades viral (m)RNA, RNaseL is activated by 2′-5′-linked oligoadenylates produced by OAS1	NS1 (IFLUAV), E3L, D9, D10 (VACV), σ3 (ReoV), Tat (HIV), ns2 (murine hepatitis virus), VP3 (RotaV), L* (Theiler's virus), hairpin RNA structure (poliovirus), genome adaptation (HCV)	Y	Y
	SAMHD1	Arteri-, pox-, herpesviruses	Hydrolyzes cellular dNTP and degrades viral RNA	None known	Y	Y
	APOBEC3 family	Herpes-, papillomaviruses	Induces hypermutation by deamination	None known	Some	Some
7	PKR	Herpes-, orthomyxo-, retro-, flavi-, reo-, adeno-, poxviruses	Inhibits mRNA translation by eIF2α phosphorylation	NS1 (IFLUAV), E2, NS5A (HCV), TRS1, IRS1 (HCMV), K3L, E3L (VACV), US11 (HSV1), vIRF-2, LANA2 (KSHV), NSs (RVFV), σ3, σ4 (ReoV), SM, EBER-1 (EBV), Tat (HIV), VAI RNAs (AV), C8L, K3L (Swinepoxv), Nsp3 (RotaV), γ(1)34.5 (HSV-1)	Y	Y
	SLFN11	HIV, other retroviruses	Inhibits viral protein synthesis by altering tRNA function	None known	Y	Y
	ZAP	Retro-, filo-, hepadna-, togaviruses	Recruits RNA exosome complex to degrade viral RNA	None known	Y	Y
	IFIT family	Flavi-, bunya-, rhabdo-, orthomyxo-, picorna-, coronaviruses	Inhibits cap- and IRES-dependent translation by binding to eIF3 (HCV), binding and degradation of PPP-RNA (RVFV, VSV, IFLUAV) and RNA lacking 2′-O methylation (WNV, JEV)	2′-O methylation of viral RNA (WNV, SARS-CoV, VACV), hairpin structures near the 5′ ends of viral RNA (VEEV), masking of the 5′ end by Vpg (EMCV)	Y	Y
8	HERC5 (+ISG15)	HIV, MLV, HPV, IFLUAV	Inhibits HIV and MLV assembly by ISGylation of Gag, ISGylation of IFLUAV NS1 and HPV L1 capsid reduces infectious virus yield	NS1 (?) (IFLUAV)	Y	Y
9	Tetherin/BST2/CD317	Retro-, flavi-, filo-, rhabdo-, herpes-, corona-, paramyxo-, arena-, toga-, hepadnaviruses	Prevents virus release by tethering budding progeny virions to the plasma membrane of the infected cell	Vpu (HIV-1 M/N, SIVgsn/mon/mus), Nef (most SIV, HIV-1 O), K5 (KSHV), Env (HIV-2, EBOV, MARV, SIVagm), Nsp1 (CHIKV), gM (HSV-1), HA/NA (pandemic IFLUAV), F/HN (SeV), HBs (HBV)	Y	Y

 DOI http://dx.doi.org/10.1016/j.cell.2015.10.019 See online version for legend and references.

SnapShot: Antiviral Restriction Factors

Silvia F. Kluge, Daniel Sauter, and Frank Kirchhoff
Institute of Molecular Virology, Ulm University Medical Center, 89081 Ulm, Germany

Restriction factors are cellular proteins that inhibit viral replication and represent a first line of defense against viral pathogens. They show an enormous structural and functional diversity and target almost every step of the viral replication cycle. Although there is no unambiguous definition of restriction factors (Doyle et al., 2015), these proteins frequently share several characteristics: they are germ-line encoded, cell-intrinsic proteins that can be found in almost all cell types. While their expression is often upregulated by interferons (IFNs), many of them are constitutively expressed, allowing them to act very early during viral infection. Restriction factors frequently target conserved viral components, such as the viral genomes or membranes, and may thus be active against diverse viral families. Notably, some of them are so-called moonlighting proteins, also exhibiting biological functions outside of immunity. In some cases, restriction of viral replication may result from a cell-regulatory function rather than direct interference with the viral replication cycle. Viruses have evolved sophisticated means to evade or directly counteract many restriction factors. As a consequence of the continuous arms race with their viral antagonists, restriction factors usually evolve rapidly and show evolutionary signatures of adaptation. Sites under positive selection often directly interact with viral components, either to target them for inhibition or because they are being targeted by viral antagonists. As a consequence of virus-host adaptation, restriction factors are usually less effective against viruses in their natural hosts but represent potent barriers against cross-species transmissions. Finally, their specific interaction with viral components allows some restriction factors to act as pattern recognition receptors that do not only directly inhibit viral pathogens, but also sense them to induce antiviral immune responses.

The term “restriction factor” was established in the early 1970s, when researchers discovered that expression of Fv1 protects mice against infection by an otherwise lethal dose of MLV (Lilly, 1970). Later, it became evident that primate lentiviruses, such as HIV-1, are subject to similar restrictions. A functional screen for suppressors of HIV-1 identified rhesus TRIM5α as a potent inhibitor and determinant of retroviral species specificity (Stremlau et al., 2004). Similar to Fv1, TRIM5α and the related TRIM-CypA protein target incoming retroviral capsids and block viral replication by preventing viral cDNA synthesis. Other well-characterized retroviral restriction factors include APOBEC3G, Tetherin, and SAMHD1. APOBEC3G is a cytidine deaminase that is packaged into viral particles and inhibits viral cDNA synthesis by affecting the processivity of reverse transcription and by causing inactivating G-to-A hypermutations in the proviral genome (Sheehy et al., 2002). Tetherin inhibits the release of budding progeny virions because one of its two membrane anchors is inserted into the viral envelope while the other remains in the cell membrane (Van Damme et al., 2008; Neil et al., 2008). SAMHD1 suppresses reverse transcription in non-dividing cells by depleting dNTPs, which are required for effective cDNA synthesis, and perhaps also by degrading viral RNA (Hrecka et al., 2011; Laguette et al., 2011). With the exception of TRIM5α, that is evaded by viral capsid mutations, these restriction factors are all counteracted by accessory proteins of HIV and related lentiviruses: APOBEC3 proteins by Vif, Tetherin by Vpu of pandemic HIV-1 group M as well as Nef of many other primate lentiviruses, and SAMHD1 by HIV-2 and SIV Vpx or Vpr proteins. Very recently, SERINC5 and SERINC3 have been identified as the enigmatic factors that impair the infectivity of HIV and SIV particles and are antagonized by the viral protein Nef (Rosa et al., 2015; Usami et al., 2015).

Cellular proteins inhibiting HIV-1 have received enormous research interest, and a variety of additional antiviral factors, such as IFITM proteins, CH25H, KAP1/TRIM28, 90K, MOV10, MxB, SLFN11, and ZAP have been described. The discovery of all of these factors has relevance far beyond HIV/AIDS and other retroviruses because many of them have broad antiviral activity. For example, Tetherin suppresses the release of a large variety of enveloped viruses, including filo-, rhabdo-, arena-, and herpesviruses. Similarly, IFITMs and CH25H may impair virion infectivity of diverse virus families by altering the lipid composition of the viral membrane. Another striking example of a broadly active antiviral protein is PKR. This kinase inhibits viral mRNA translation by inhibiting the initiation factor eIF2α.

The definition of a “real” restriction factor is intensively debated. Viruses are interacting with and hijacking hundreds of cellular proteins to ensure efficient viral replication. Thus, overexpression or knockdown of many cellular factors may result in the identification of proteins with putative antiviral effects. Moreover, only a minority of the antiviral factors described to date show all features reported to be characteristic for a restriction factor. In fact, antiviral proteins without any (known) viral antagonist or evasion mechanism (e.g., IFITMs and SLFN11) have been proposed to be called “resistance factors” (Doyle et al., 2015). Here, we more broadly apply the term “restriction factor” to intrinsic cellular factors known to display antiviral activity. We apologize to both the purists who apply criteria that are more stringent and to all of the scientists who discovered interesting antiviral factors that we did not mention. We are only just beginning to understand the enormous diversity of antiviral factors and the highly sophisticated ways exploited by viruses to antagonize or evade them. No matter which definition of a restriction factor we apply, there will certainly be discoveries of novel antiviral proteins that will not satisfy the criteria.

ABBREVIATIONS

Antiviral factors: IFITM, interferon-induced transmembrane protein; SERINC, serine incorporator; CH25H, cholesterol 25-hydroxylase; TRIM, tripartite motif-containing protein; Fv1, Friend virus susceptibility-1; APOBEC3, apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3; SAMHD1, SAM domain and HD domain-containing protein 1; MxA, myxovirus resistance gene A; MxB, myxovirus resistance gene B; KAP1, KRAB-associated protein 1; RSAD2, radical S-adenosyl methionine domain-containing 2; IFI16, interferon-inducible protein 16; OAS1, 2’-5’-oligoadenylate synthetase 1; PKR, (ds)RNA-dependent protein kinase R; SLFN11, Schlafen family member 11; ZAP, zinc-finger antiviral protein; IFIT, interferon-induced protein with tetratricopeptide repeats; HERC5, HECT and RLD domain-containing E3 ubiquitin protein ligase 5; ISG15, interferon-stimulated gene 15; BST2, bone marrow stromal cell antigen 2.

Viruses: HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; MLV, murine leukemia virus; EIAV, equine infectious anemia virus; IFLUAV, influenza A virus; THOV, Thogoto virus; HCV, hepatitis C virus; DENV, Dengue virus; HPV, human papilloma virus; HCMV, human cytomegalovirus; HSV1, herpes simplex virus 1; LACV, La Crosse encephalitis virus; RVFV, Rift Valley fever virus; BUNV, bunyamweravirus; VSV, vesicular stomatitis virus; WNV, West Nile virus; JEV, Japanese encephalitis virus; KSHV, Kaposi’s sarcoma-associated herpesvirus; VACV, vaccinia virus; reoV, reovirus; EBV, Epstein-Barr virus; RotaV, rotavirus; AV, adenovirus; VEEV, Venezuelan equine encephalitis virus; SARS-CoV, severe acute respiratory syndrome corona virus; EMCV, encephalomyocarditis virus; MARV, Marburg virus; CHIKV, Chikungunya virus; SeV, Sendai virus; HBV, hepatitis B virus; EBOV, Ebola virus.

Viral proteins: Nef, negative factor; NS5A, nonstructural protein 5A; Vif, viral infectivity factor; Vpr, viral protein R; Vpu, viral protein unknown; Vpx, viral protein X; Env, envelope; vIRF-2, viral IRF2-like protein; US11, tegument protein unique short 11.

Other: eIF2α, eukaryotic translation initiation factor 2α; eIF3, eukaryotic translation initiation factor 3.

REFERENCES

Doyle, T., Goujon, C., and Malim, M.H. (2015). Nat. Rev. Microbiol. *13*, 403–413.

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S.K., Kesik-Brodacka, M., Srivastava, S., Florens, L., Washburn, M.P., and Skowronski, J. (2011). Nature *474*, 658–661.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., Yatim, A., Emiliani, S., Schwartz, O., and Benkirane, M. (2011). Nature *474*, 654–657.

Lilly, F. (1970). J. Natl. Cancer Inst. *45*, 163–169.

Neil, S.J.D., Zang, T., and Bieniasz, P.D. (2008). Nature *451*, 425–430.

Rosa, A., Chande, A., Ziglio, S., De Sanctis, V., Bertorelli, R., Goh, S.L., McCauley, S.M., Nowosielska, A., Antonarakis, S.E., Luban, J., et al. (2015). Nature. Published online September 30, 2015. http://dx.doi.org/ 10.1038/nature15399.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Nature *418*, 646–650.

Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., and Sodroski, J. (2004). Nature *427*, 848–853.

Usami, Y., Wu, Y., and Gottlinger, H.G. (2015) Nature. Published online September 30, 2015. http://dx.doi.org/ 10.1038/nature15400.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R.L., Mitchell, R., Johnson, M.C., Stephens, E.B., and Guatelli, J. (2008). Cell Host Microbe *3*, 245–252.

  DOI http://dx.doi.org/10.1016/j.cell.2015.10.019